CSCI 5541: Natural Language Processing

Lecture 14: All about Data, Annotation, and Evaluation

Dongyeop Kang (DK), University of Minnesota

dongyeop@umn.edu | twitter.com/dongyeopkang | dykang.github.io

computer science
& engineering

MINNESOTA NLP EST. 2021

University of Minnesota
Driven to Discover®

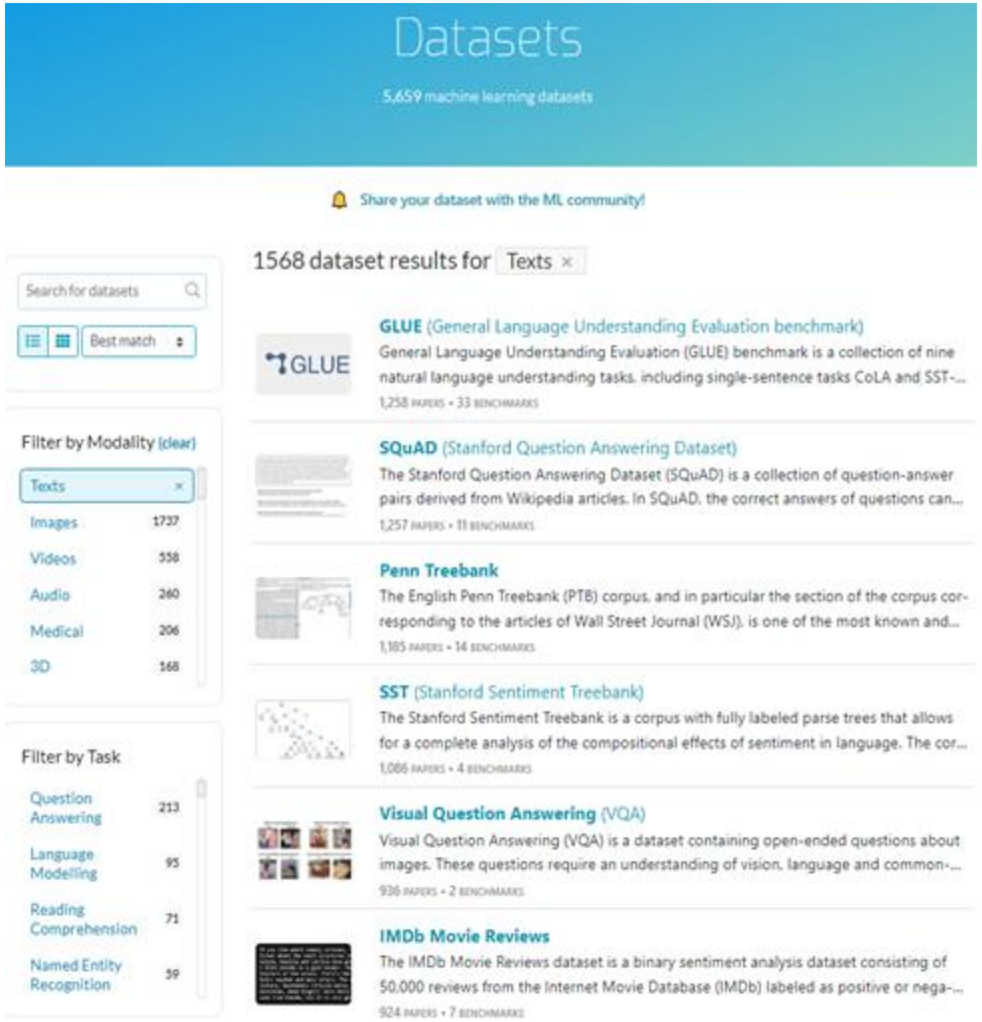
Outline

- ❑ Annotation terms, examples, and process
- ❑ Qualitative coding
- ❑ Recruiting annotators (coders)
- ❑ Annotation quality assessment
- ❑ Annotation tools
- ❑ Issues in annotation
- ❑ Advanced annotation techniques
- ❑ LLMs as Annotators and Synthetic Data

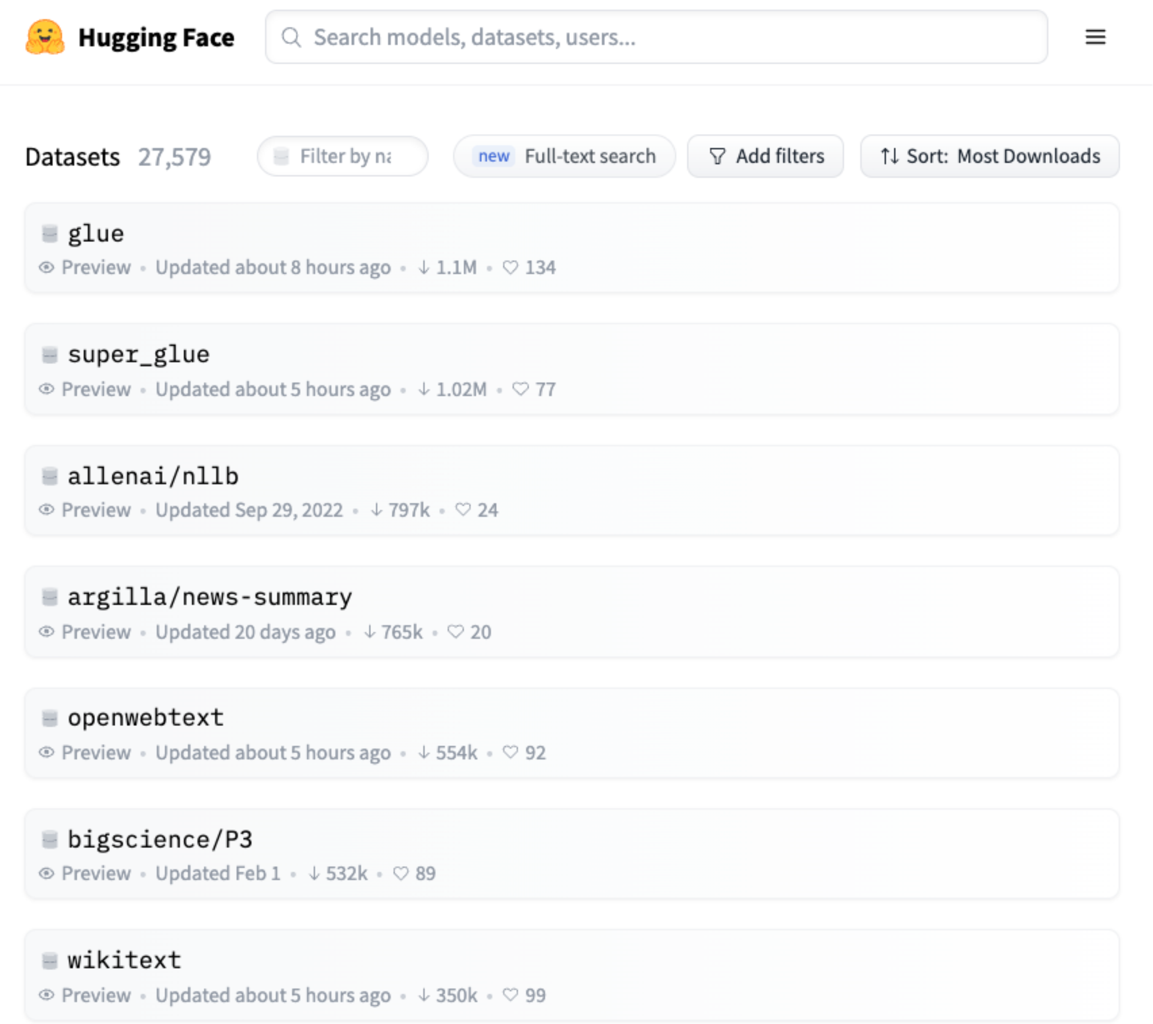
Annotation

- Despite the emergent ability of LLMs, fine-tuned models trained on annotated dataset still shows better performance.
- High-quality data means high-performance algorithms
- Just providing large amounts of data doesn't help the model understand and learn to speak. The data needs to be guided in such a way that the computer can more easily find patterns and inferences.
- Any metadata (e.g., tags, structures, categories, orders) used to mark up elements of the dataset is called annotation.
- But, in order for the algorithms to learn efficiently and effectively, the annotation must be accurate, and relevant to the task the machine is being asked to perform.

https://paperswithcode.com/datasets

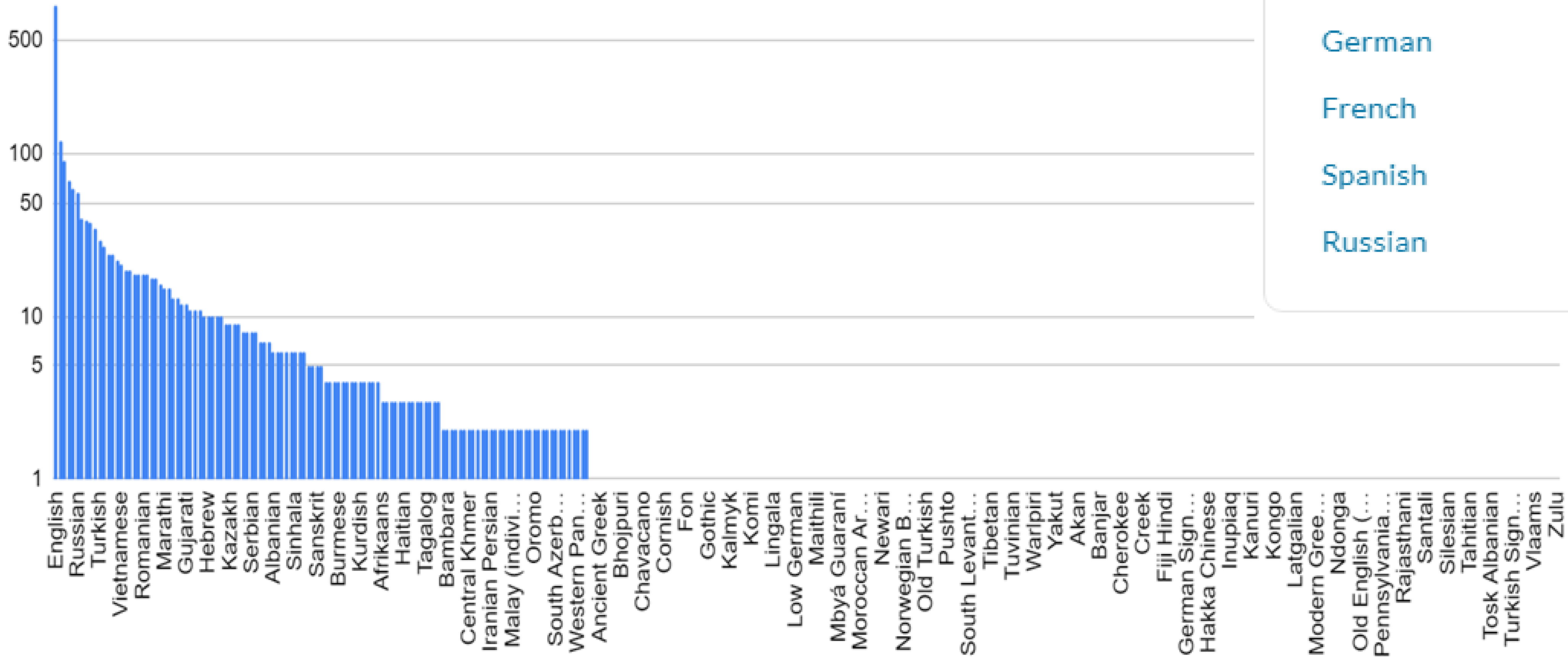


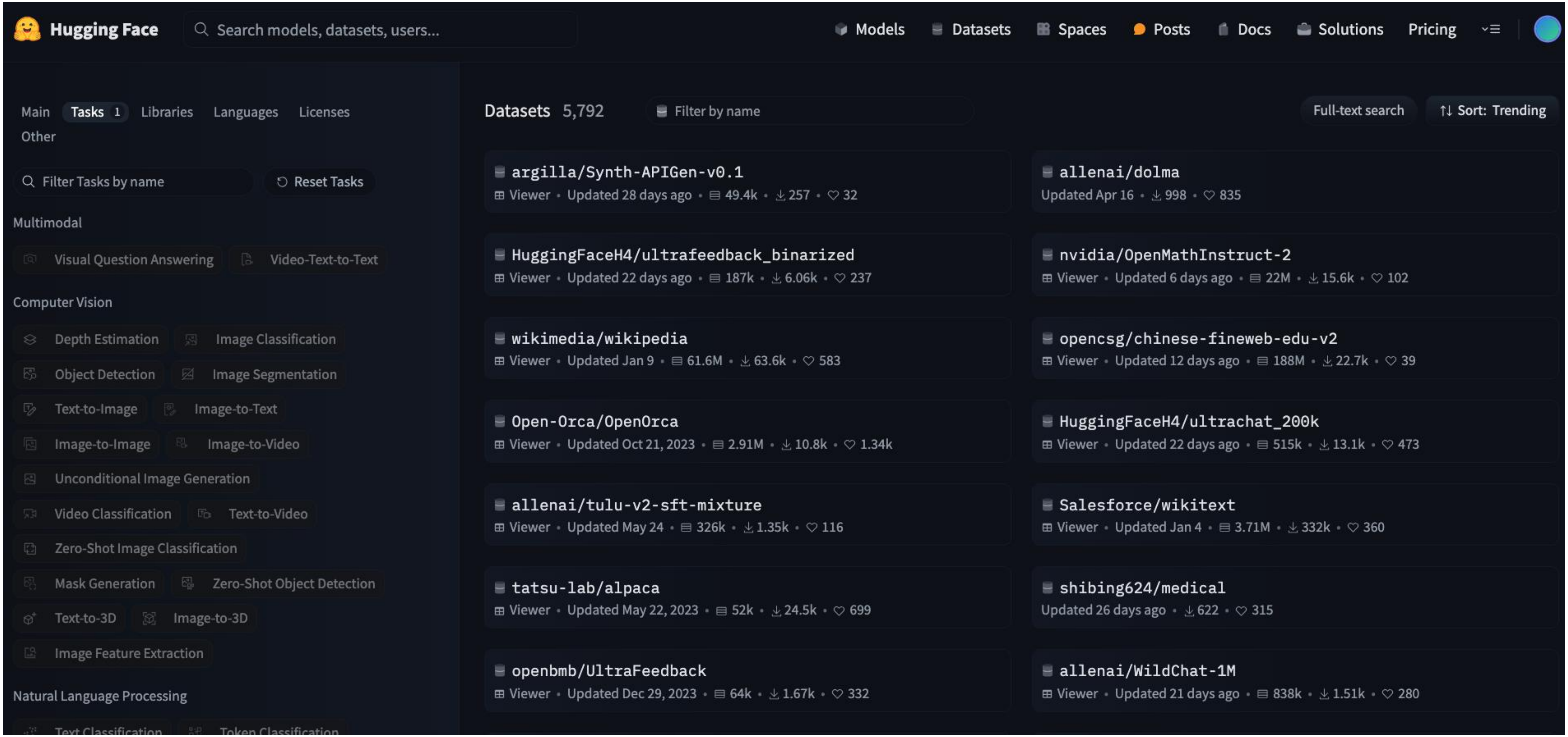
https://huggingface.co/datasets?sort=downloads



https://paperswithcode.com/datasets

Current benchmark datasets are skewed to high-resource languages

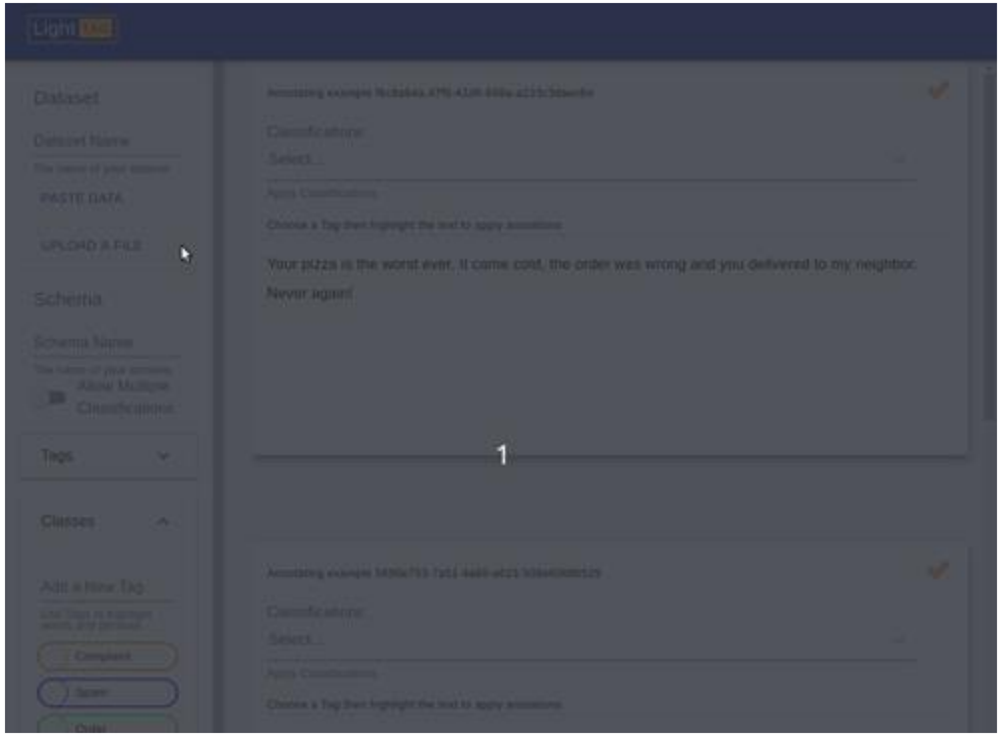



Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Solutions
Pricing
Main
Tasks 1
Libraries
Languages
Licenses
Other
Filter Tasks by name
Reset Tasks
Multimodal
Visual Question Answering
Video-Text-to-Text
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Text-to-Image
Image-to-Text
Image-to-Image
Image-to-Video
Unconditional Image Generation
Video Classification
Text-to-Video
Zero-Shot Image Classification
Mask Generation
Zero-Shot Object Detection
Text-to-3D
Image-to-3D
Image Feature Extraction
Natural Language Processing
Datasets 5,792
Filter by name
Full-text search
Sort: Trending
argilla/Synth-APIGen-v0.1
Viewer • Updated 28 days ago • 49.4k • 257 • 32
allenai/dolma
Updated Apr 16 • 998 • 835
HuggingFaceH4/ultrafeedback_binarized
Viewer • Updated 22 days ago • 187k • 6.06k • 237
nvidia/OpenMathInstruct-2
Viewer • Updated 6 days ago • 22M • 15.6k • 102
wikimedia/wikipedia
Viewer • Updated Jan 9 • 61.6M • 63.6k • 583
opencsg/chinese-fineweb-edu-v2
Viewer • Updated 12 days ago • 188M • 22.7k • 39
Open-Orca/OpenOrca
Viewer • Updated Oct 21, 2023 • 2.91M • 10.8k • 1.34k
HuggingFaceH4/ultrachat_200k
Viewer • Updated 22 days ago • 515k • 13.1k • 473
allenai/tulu-v2-sft-mixture
Viewer • Updated May 24 • 326k • 1.35k • 116
Salesforce/wikitext
Viewer • Updated Jan 4 • 3.71M • 332k • 360
tatsu-lab/alpaca
Viewer • Updated May 22, 2023 • 52k • 24.5k • 699
shibing624/medical
Updated 26 days ago • 622 • 315
openbmb/UltraFeedback
Viewer • Updated Dec 29, 2023 • 64k • 1.67k • 332
allenai/WildChat-1M
Viewer • Updated 21 days ago • 838k • 1.51k • 280

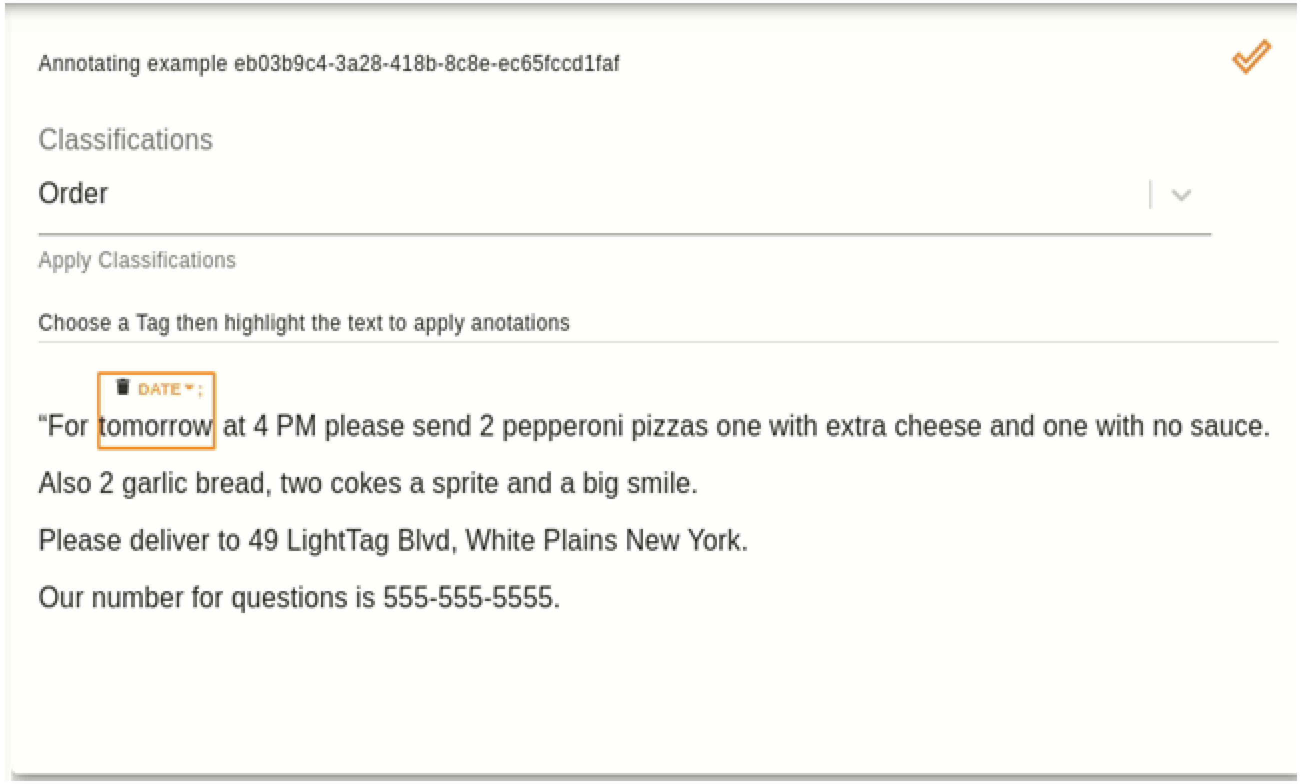
Terms

- Datasets of natural language are referred to as corpora
- A single set of data annotated with the same specification is called an annotated corpus.
- A dataset is a collection of examples that need to be annotated.
  - A class is a particular classification option.
    - ✓ E.g., Positive or Negative and email can be Spam or Ham.
  - A tag is a description name for an entity type.
    - ✓ E.g., Person (Jane), Country (Madagascar), Topping (Pepperoni) and Emotion (Fascinated).
  - A response to particular question or prompt
    - ✓ E.g., "the answer is 4"
- A schema
  - Everyone to use the same collection of tags and classes or
  - Pick and choose their own tags and classes.

Types of annotations

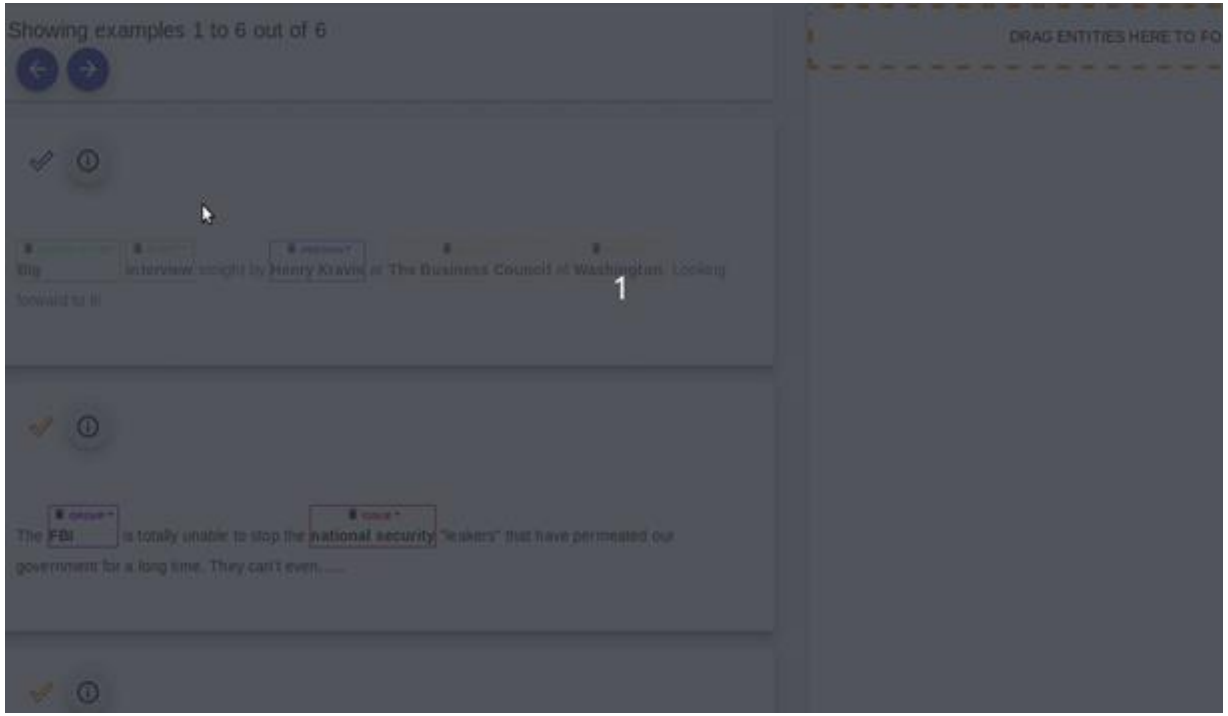


Document classification

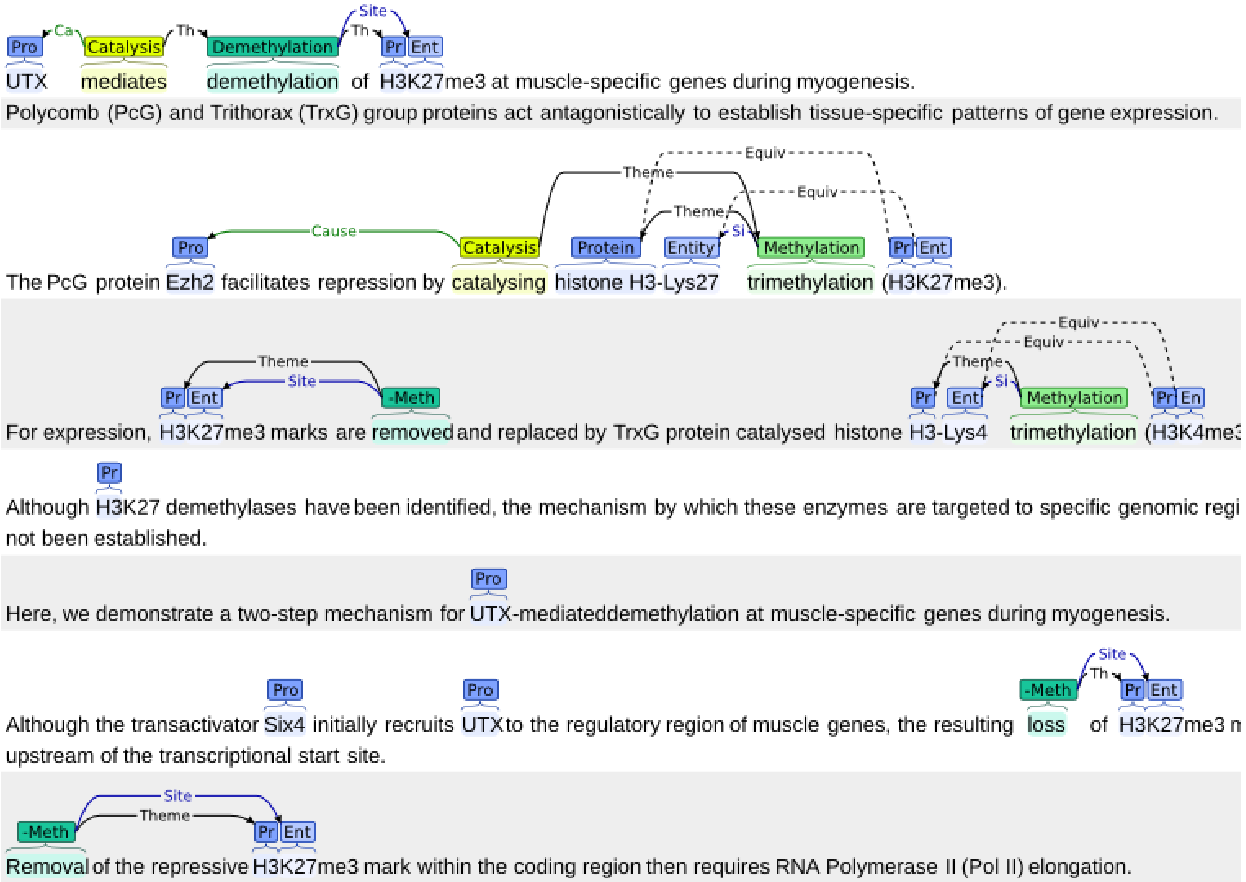


Entity annotation

Types of annotations

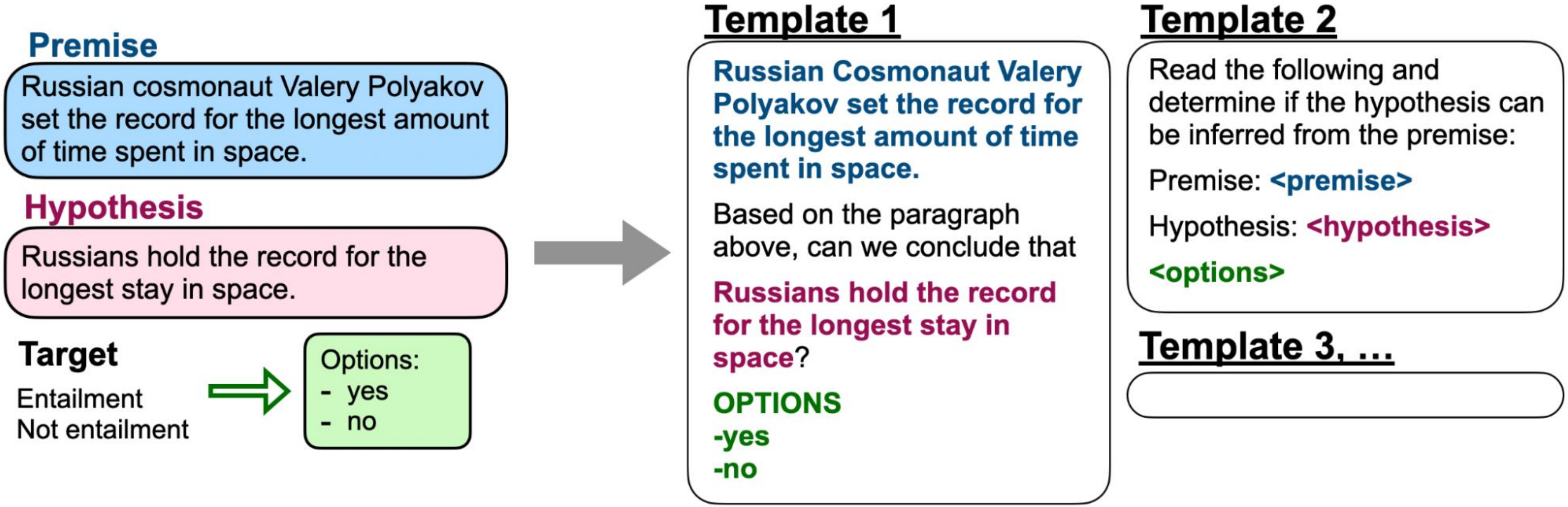


Relation annotation



Discourse relation annotation

Types of annotations

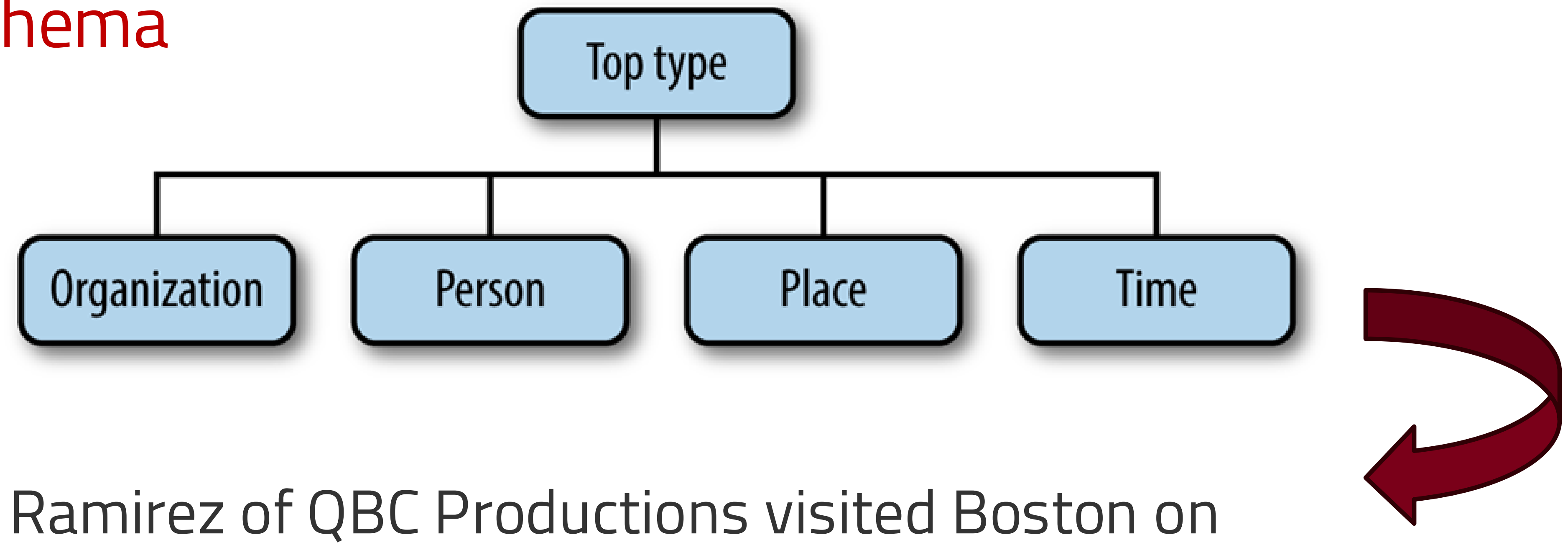


Questions for collecting the ideal dataset?

- What is the target accuracy you are looking for?
- Can it be achieved it by better models or more data?
  - How many annotations are enough to ensure high accuracies?
- How representative is your dataset?
  - domain vocabulary, format, genre of the text, etc
- Is your dataset balanced, containing instances of each class?
- How clean is your dataset?

Examples on semantic types/role labeling

Schema



Ms. Ramirez of QBC Productions visited Boston on Saturday, where she had lunch with Mr. Harris of STU Enterprises at 1:15 pm.

Semantic Types

[Ms. Ramirez]$_{\text{Person}}$ of [QBC Productions]$_{\text{Organization}}$ visited [Boston]$_{\text{Place}}$ on [Saturday]$_{\text{Time}}$, where she had lunch with [Mr. Harris]$_{\text{Person}}$ of [STU Enterprises]$_{\text{Organization}}$ at [1:15 pm]$_{\text{Time}}$.

Semantic Role Labeling

- Basics for Question Answering,
  - the who, what, where, and when of a sentence.

Agent	The event participant that is doing or causing the event to occur
Theme/figure	The event participant who undergoes a change in position or state
Experiencer	The event participant who experiences or perceives something
Source	The location or place from which the motion begins; the person from whom the theme is given
Goal	The location or place to which the motion is directed or terminates
Recipient	The person who comes into possession of the theme
Patient	The event participant who is affected by the event
Instrument	The event participant used by the agent to do or cause the event
Location/ground	The location or place associated with the event itself

The man painted the wall with a paint brush.

Mary walked to the café from her house.

John gave his mother a necklace.

My brother lives in Milwaukee.

[The man]$_{agent}$ painted [the wall]$_{patient}$ with [a paint brush]$_{instrument}$.

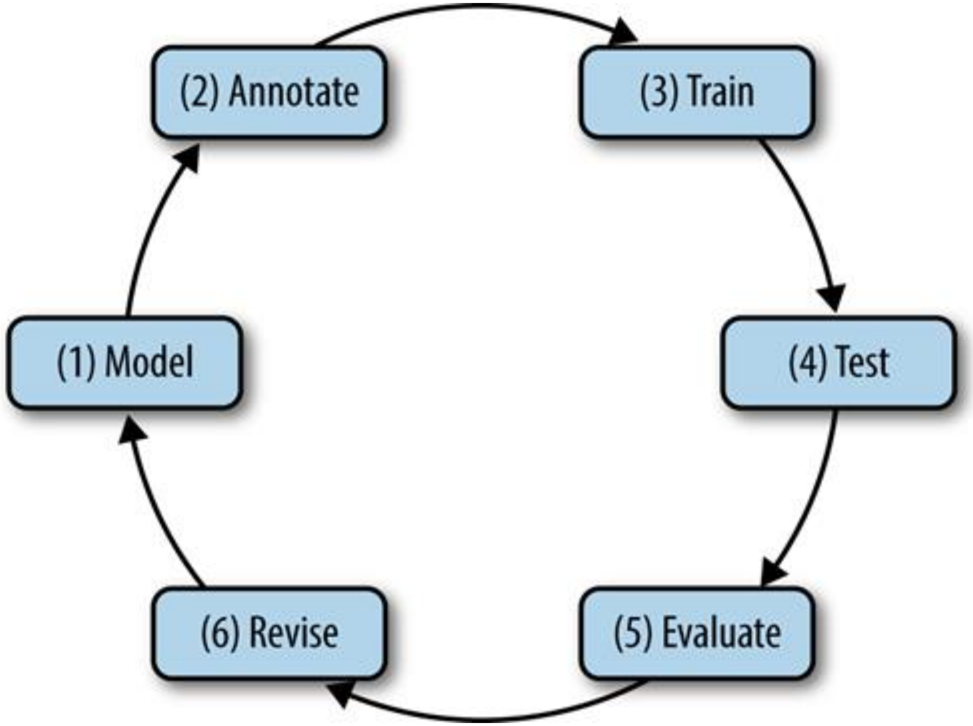
[Mary]$_{figure}$ walked to [the cafe]$_{goal}$ from [her house]$_{source}$.

[John]$_{agent}$ gave [his mother]$_{recipient}$ [a necklace]$_{theme}$.

[My brother]$_{theme}$ lives in [Milwaukee]$_{location}$.

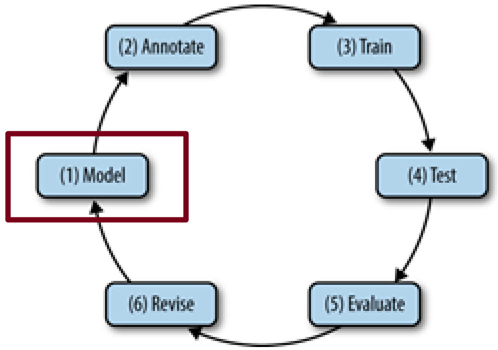
Annotation process

Annotation Development Cycle



MATTER methodology (Pustejovsky 2006)

Model the Phenomenon



A model, M, can be seen as a triple, M = <T,R,I>.

- ❑ A vocabulary of terms, T,
- ❑ The relations between these terms, R,
- ❑ Their interpretation, I.

Terms = {Document_type, Spam, Not-Spam}

Relations = {Document_type ::= Spam | Not-Spam}

Interpretation = { Spam = "something we don't want!",

Not-Spam = "something we do want!"}

Terms = {Named_Entity, Organization, Person, Place, Time}

Relations = {Named_Entity ::= Organization | Person | Place | Time}

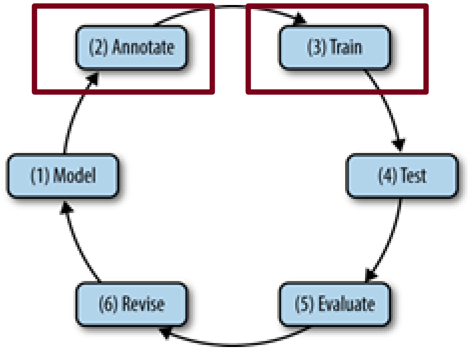
Interpretation = { Organization = "list of organizations in a database",

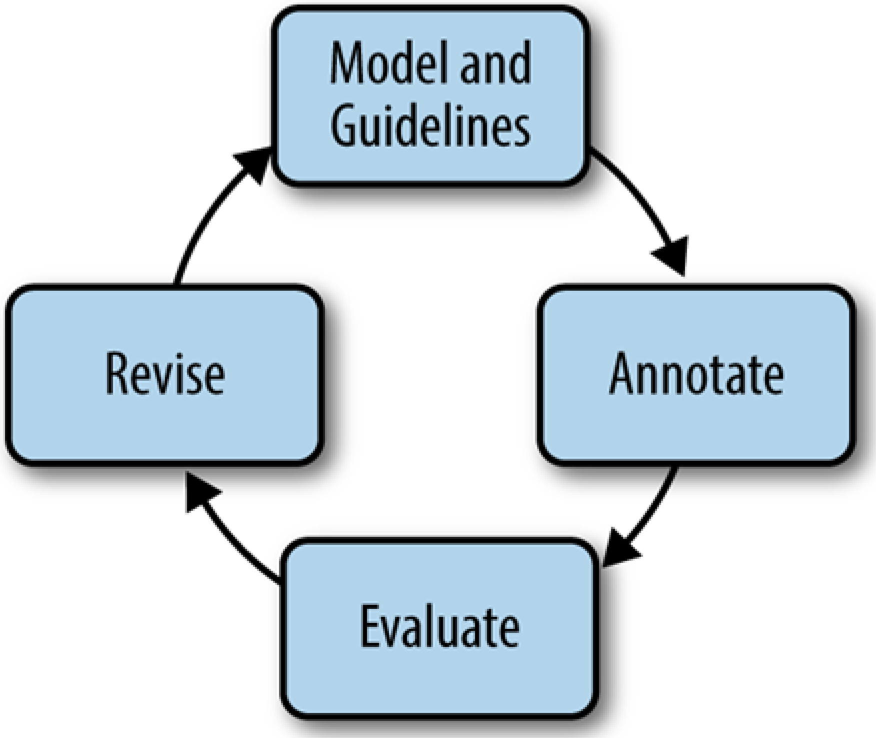
Person = "list of people in a database",

Place = "list of countries, geographic locations, etc.",

Time = "all possible dates on the calendar"}

Annotate with the Specification

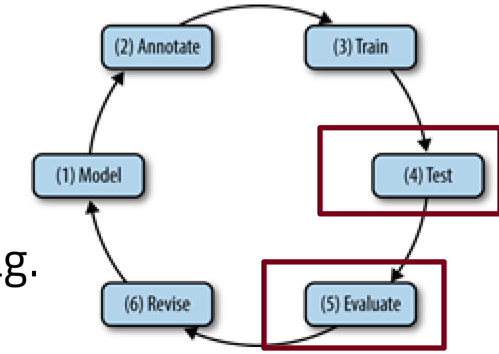




MAMA (Model-Annotate-Model-Annotate) cycle, or the “babeling” phase of MATTER.

Given the specification document encoding the model phenomenon, now you will need to train human annotators to mark up the dataset according to the tags that are important to you.

Consistency



the most problematic when comparing annotations: namely, the extent or the span of the tag.

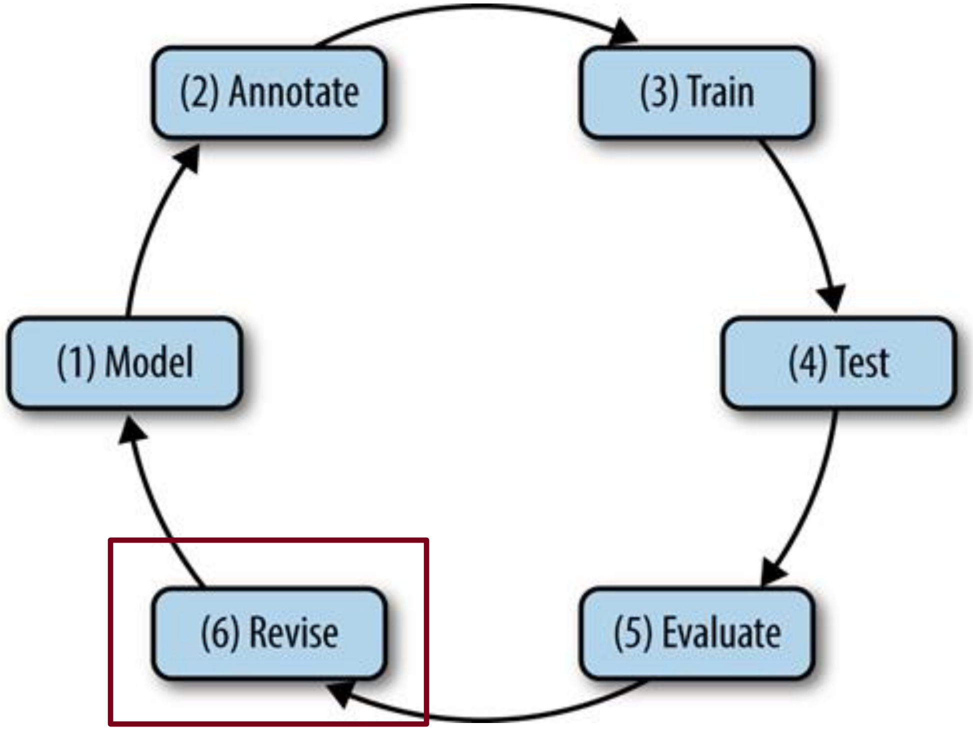
QBC Productions Inc. of East Anglia Organization

[QBC Productions]$_{\text{Organization}}$ Inc. of East Anglia

[QBC Productions Inc.]$_{\text{Organization}}$ of East Anglia

[QBC Productions Inc. of East Anglia]$_{\text{Organization}}$

Annotation Development Cycle

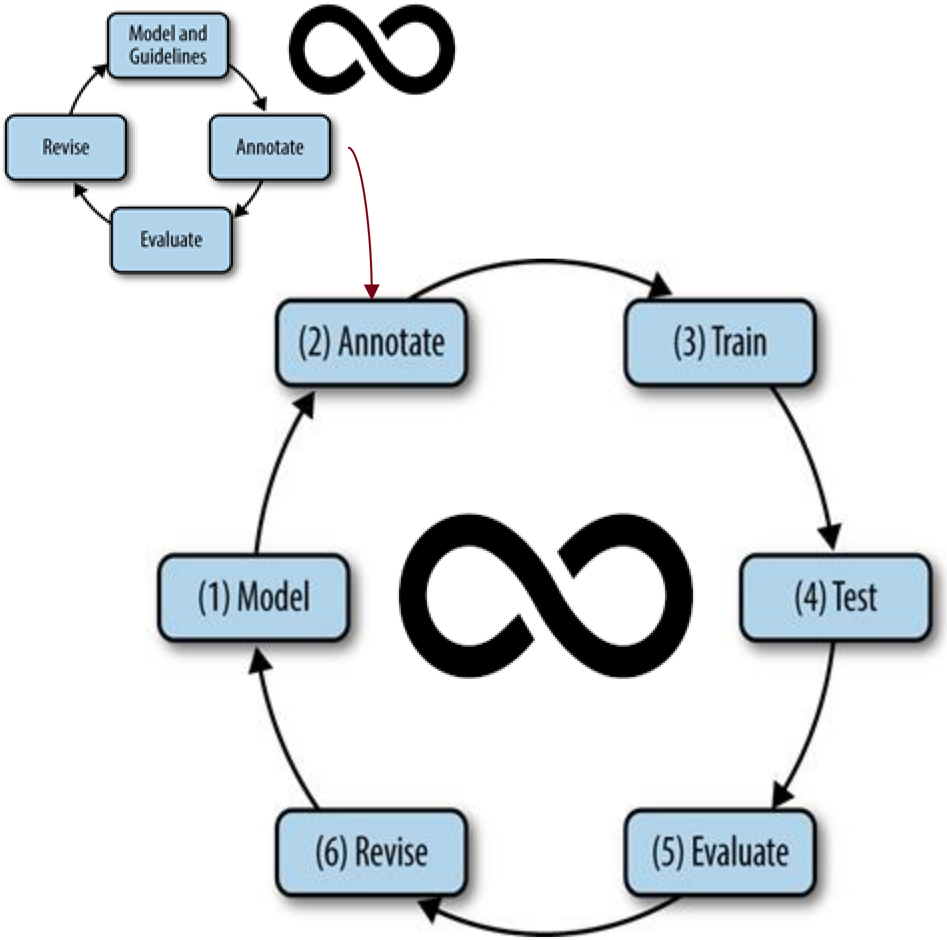


MATTER methodology (Pustejovsky 2006)

Revise

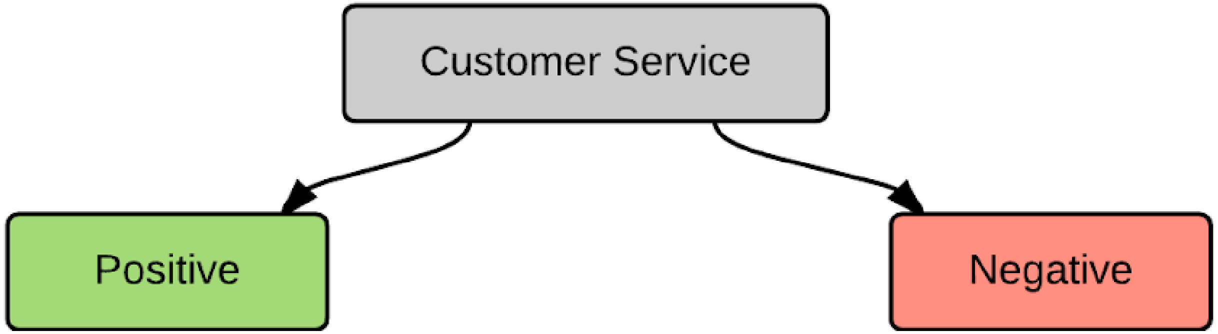
The model and the annotation specification are revisited in order to make the annotation more robust and reliable with use in the algorithm.

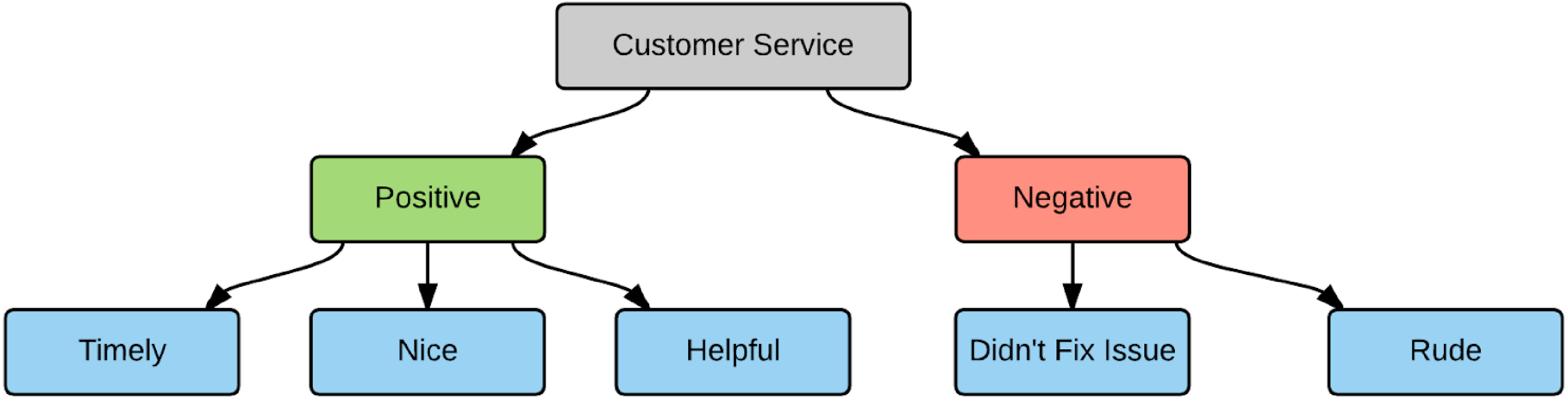
In Practice



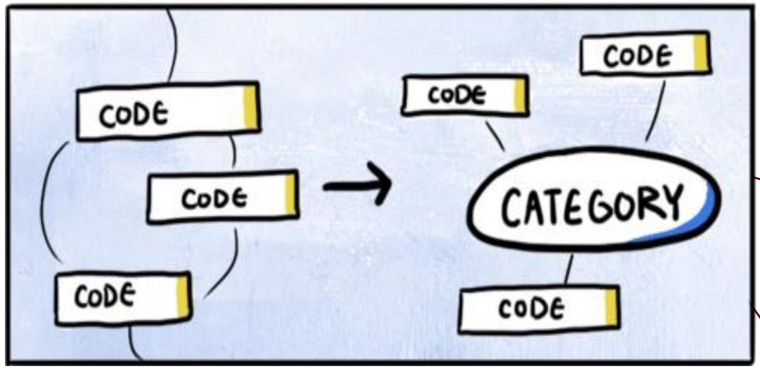
- An iterative process until you reach to the target performance
- As model performance converges, you will face edge cases in the long tail. Analyzing the long-tail and updating the schema are painful and time-consuming, but most important in practice.
- There is no single magic deep learning solution in real-world tasks; If so, your task is relatively easy or narrowed down to a very specific scope

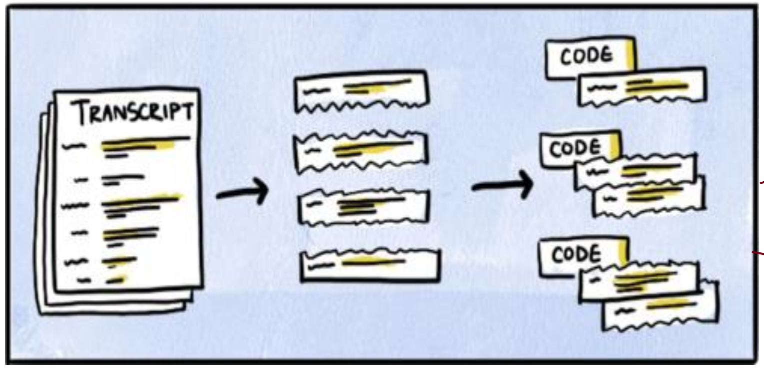
Qualitative coding

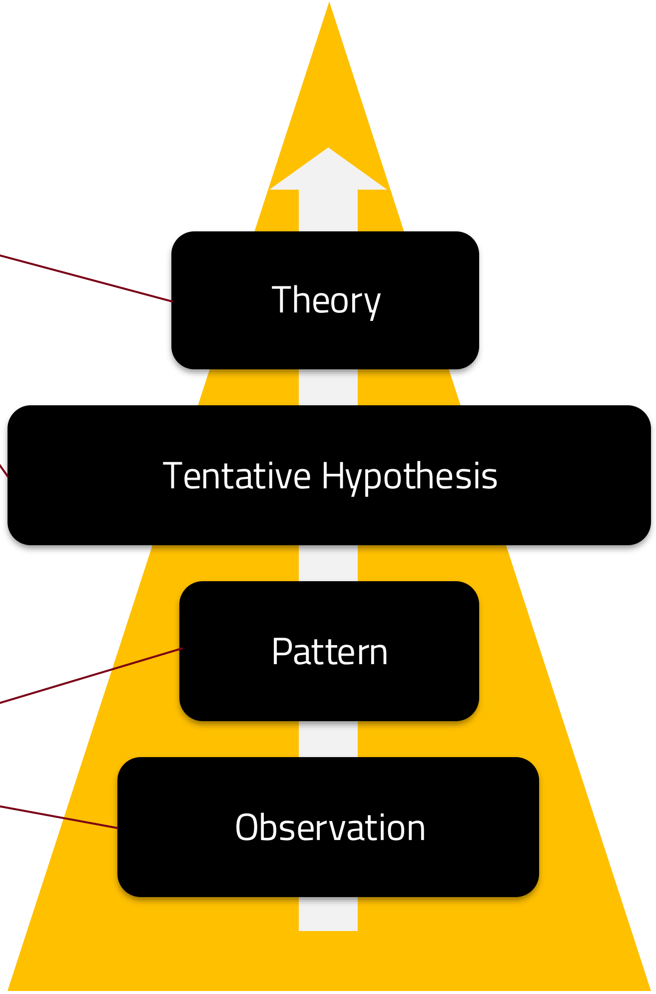

Customer Service
Positive
Negative

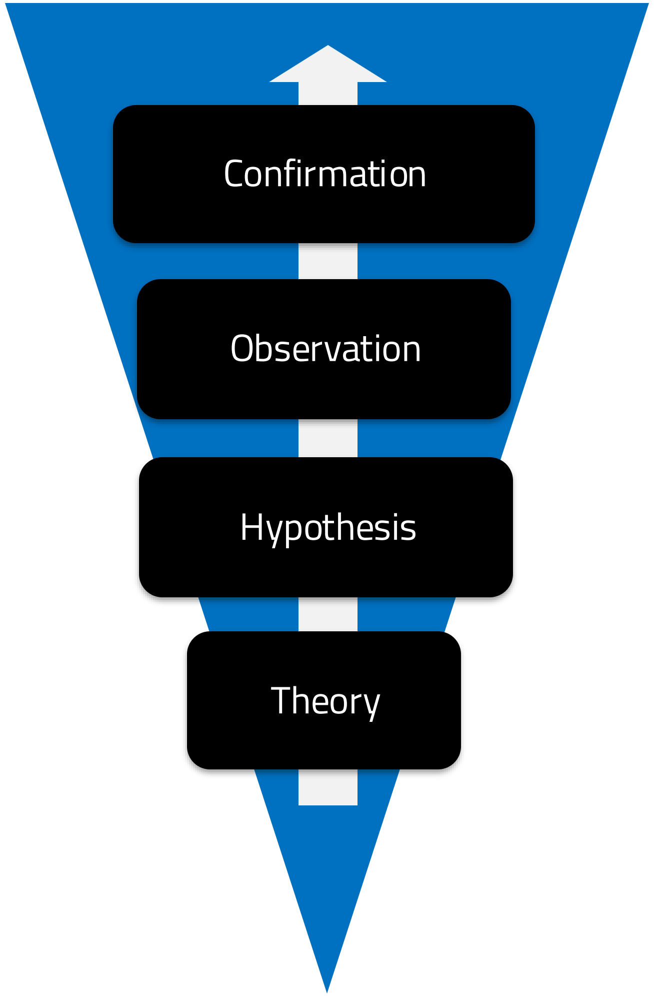

Customer Service
Positive
Negative
Timely
Nice
Helpful
Didn't Fix Issue
Rude

Deductive coding







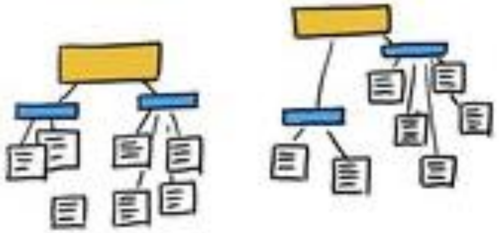


Inductive coding

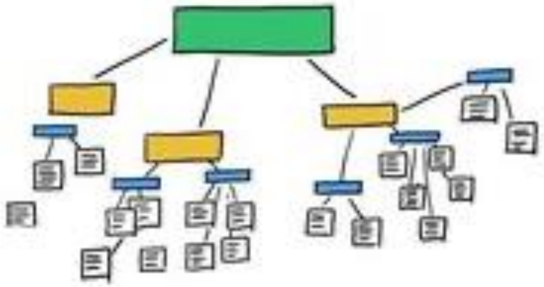
Steps in inductive coding

Open coding:
Compare snippets with snippets and create codes that connect them.

Axial coding:
Compare codes with codes and create categories (or axes) that connect them.



Selective coding:
Compare categories with categories and create the core category that connect them.



Human-AI Collaborative Taxonomy Construction

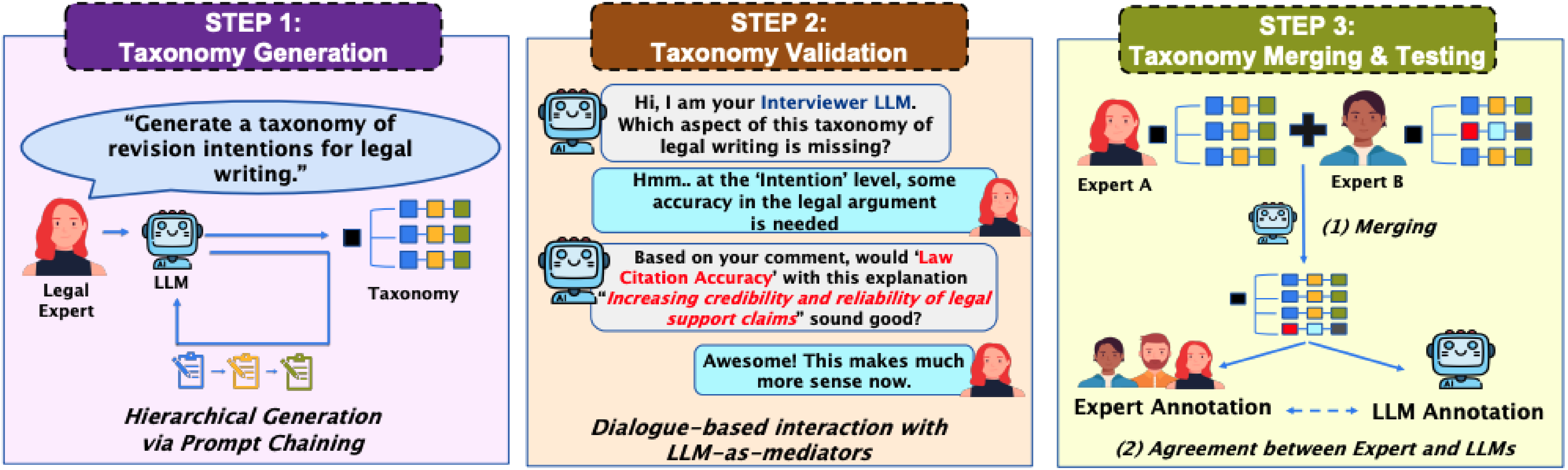
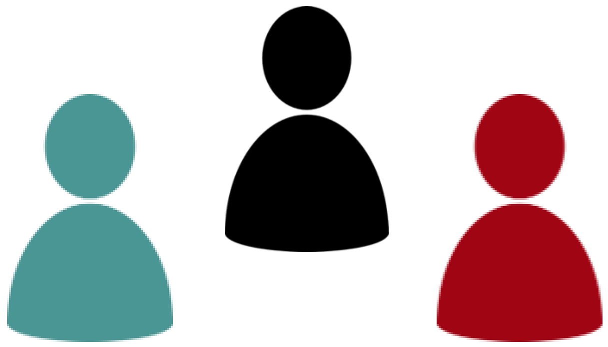


Figure 1: An end-to-end pipeline of our three-step Human-AI collaborative taxonomy construction process. For each step, we portray several design implications for better human-AI interaction strategies that were described in Section 3.

Human-AI Collaborative Taxonomy Construction: A Case Study in Profession-Specific Writing Assistants
Minhwa Lee, Zae Myung Kim, Vivek Khetan, Dongyeop Kang, In2Writing @ CHI 2024

Recruiting annotators (coders)

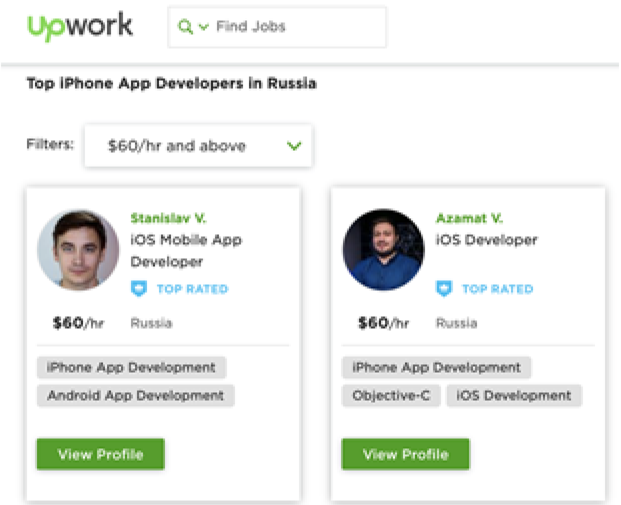


Outsourcing

- Finding capable annotators can be a tremendous headache.
- From testing, onboarding, and ensuring tax compliance to distributing, managing, and assessing the quality of projects, there's an enormous amount of hidden labor involved in annotating.

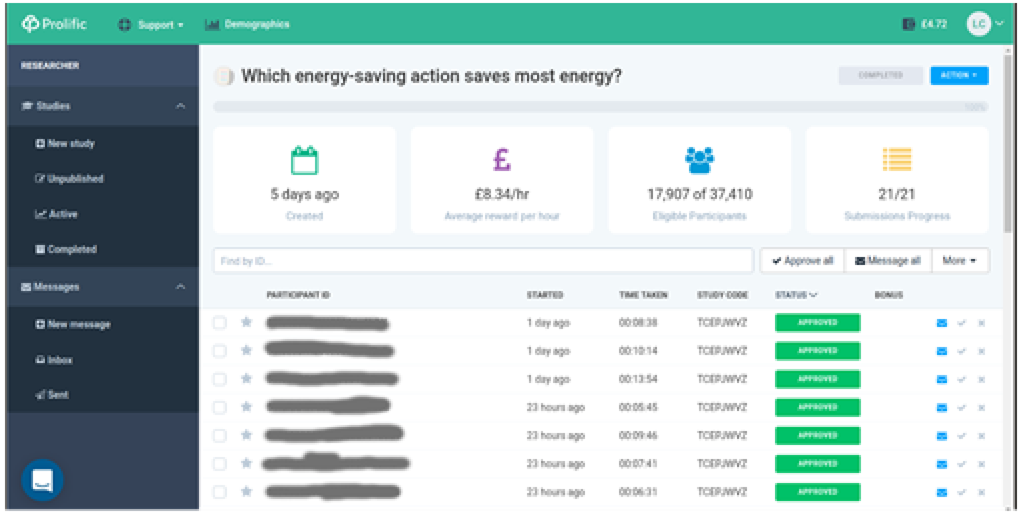
Amazon Mechanical Turk.

Best for finding people to help complete crowdsourced tasks



UpWork

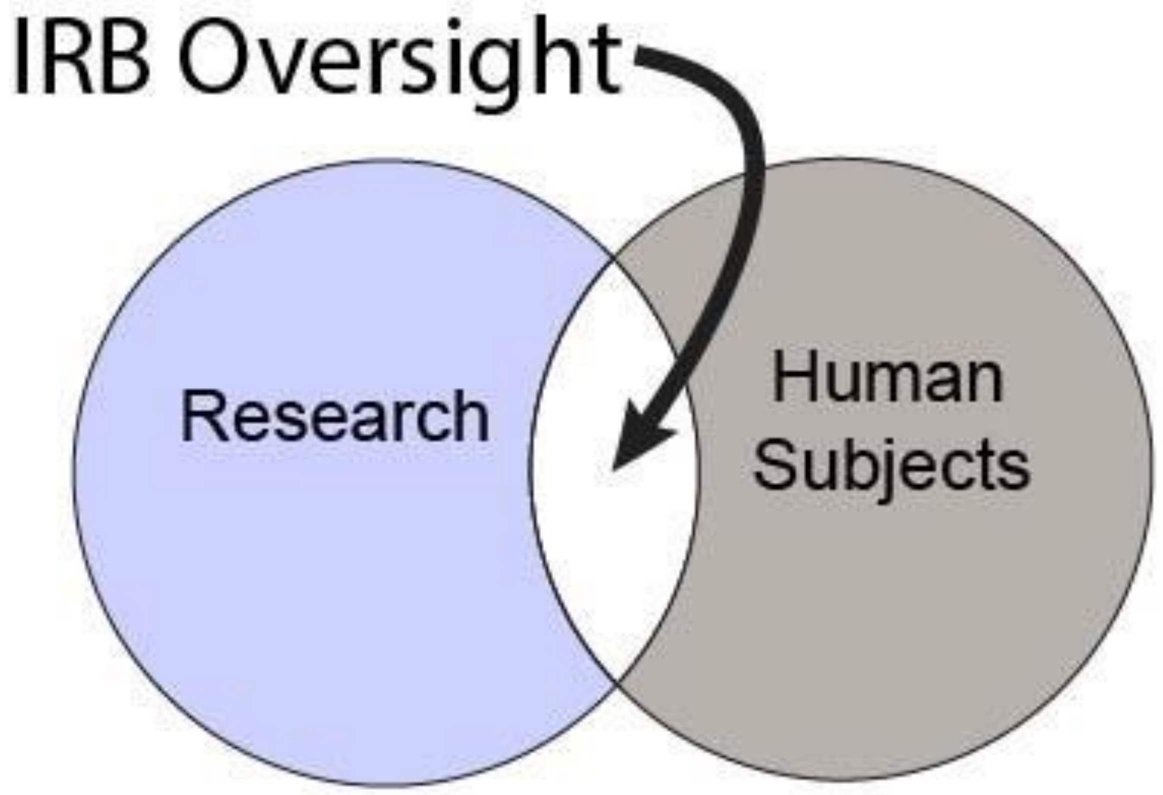
Best for finding the right freelancers to complete tasks



Prolific

Quickly find research participants you can trust.

Undergraduate students



An institutional review board (IRB) .. is a type of committee that applies research ethics by reviewing the methods proposed for research to ensure that they are ethical.

- Takes at least two months to get approval
- Before approval, you can't collect any human-subject data in your project

Annotation quality assessment

Correctness of annotations

Sentence	Coder 1	Coder 2	Agreement
We address the problem of ...... recognition	I	P	✗
Our aim is to ...recognize [x] from [y].	P	P	✓
[A] is set up as prior information, and its pose is determined by three parameters, which are [j,k and l].	M	M	✓
An efficient local gradient-based method is proposed to ..., which is combined into ... framework to estimate [V and W] by iterative evolution	P	R	✗
It is shown that the local gradient-based method can evaluate accurately and efficiently [V and W] .	R	R	✓

Observed agreement between coder 1 and 2: 60%

Inter-annotator agreement (IAA)

the probability that the raters could have agreed purely by chance.

- Relative agreement is 60% in the previous example, but chance agreement is 20%. Agreement measures need to be corrected for change agreement (Carletta, 1996)

- Kappa coefficient (Cohen 1960)
  - 1 (agreement), 0 (no correlation), -1 (disagreement)

Corrected measure:

$$K = \frac{\boxed{P(A)} - \boxed{P(E)}}{1 - \boxed{P(E)}} = \frac{0.6 - 0.2}{1 - 0.2} = 0.5$$

Step 1: Calculate relative agreement (po) between raters.

Rater 2

Rater 1

	Yes	No
Yes	25	10
No	15	20

p_o = (Both said Yes + Both said No) / (Total Ratings)

= (25 + 20) / (70) = 0.6429

the probability that the raters could have agreed purely by chance.

Step 2: Calculate the hypothetical probability of chance agreement (p_e) between raters.

		Rater 2	
		Yes	No
Rater 1	Yes	25	10
	No	15	20

$\boxed{P(\text{"Yes"}) = ((25+10)/70) * ((25+15)/70) = 0.285714}$

$P(\text{"No"}) = ((15+20)/70) * ((10+20)/70) = 0.214285$

$p_e = 0.285714 + 0.214285 = 0.5$

the probability that the raters could have agreed purely by chance.

Step 2: Calculate the hypothetical probability of chance agreement (p_e) between raters.

		Rater 2	
		Yes	No
Rater 1	Yes	25	10
	No	15	20

P ("Yes") = ((25+10)/70) * ((25+15)/70) = 0.285714

P ("No") = ((15+20)/70) * ((10+20)/70) = 0.214285

p_e = 0.285714 + 0.214285 = 0.5

Step 3: Calculate Cohen's Kappa

Rater 2

Rater 1		Yes	No
	Yes	25	10
	No	15	20

$$
\begin{aligned}
k &= (p_o - p_e) / (1 - p_e) \\
&= (0.6429 - 0.5) / (1 - 0.5) \\
&= 0.2857
\end{aligned}
$$

Interpretation of Cohen's Kappa

Value Range	Cohen's Interpretation
Below 0.20	None to slight agreement
.21–.39	Fair agreement
.40–.59	Moderate agreement
.60–.79	Substantial agreement
.80–.90	Almost perfect agreement
Above .90	Almost perfect agreement

Types of Data

Quantitative

Data that can be measured with numbers, such as duration or speed

- **Discrete**

  Whole numbers that can't be broken down, such as a number of items

- **Continuous**

  Numbers that can be broken down, such as height or weight

  - **Interval**

    Numbers with known differences between variables, such as time

  - **Ratio**

    Numbers that have measurable intervals where difference can be determined, such as height or weight

Qualitative

Non-numerical data that is categorical, such as yes/no responses or eye colour

- **Nominal**

  Data used for naming variables, such as hair colour

- **Ordinal**

  Data used to describe the order of values, such as 1 = happy, 2 = neutral, 3 = unhappy

Other IAA measures by types and their interpretation

Comparison of IRR indices in presence of research limitations					
IRR	Data	Missing Data	Number of Raters	The effect of 'chance' in agreement is minimized?	**General agreement on the significance of a numeric result?**
Cohen's Kappa	Nominal	No	2	No *	**No**
Fleiss's Kappa	Nominal	No	2≥	No *	**No**
Krippendorff's Alpha	All Data	Yes	2≥	Yes	**Yes** **

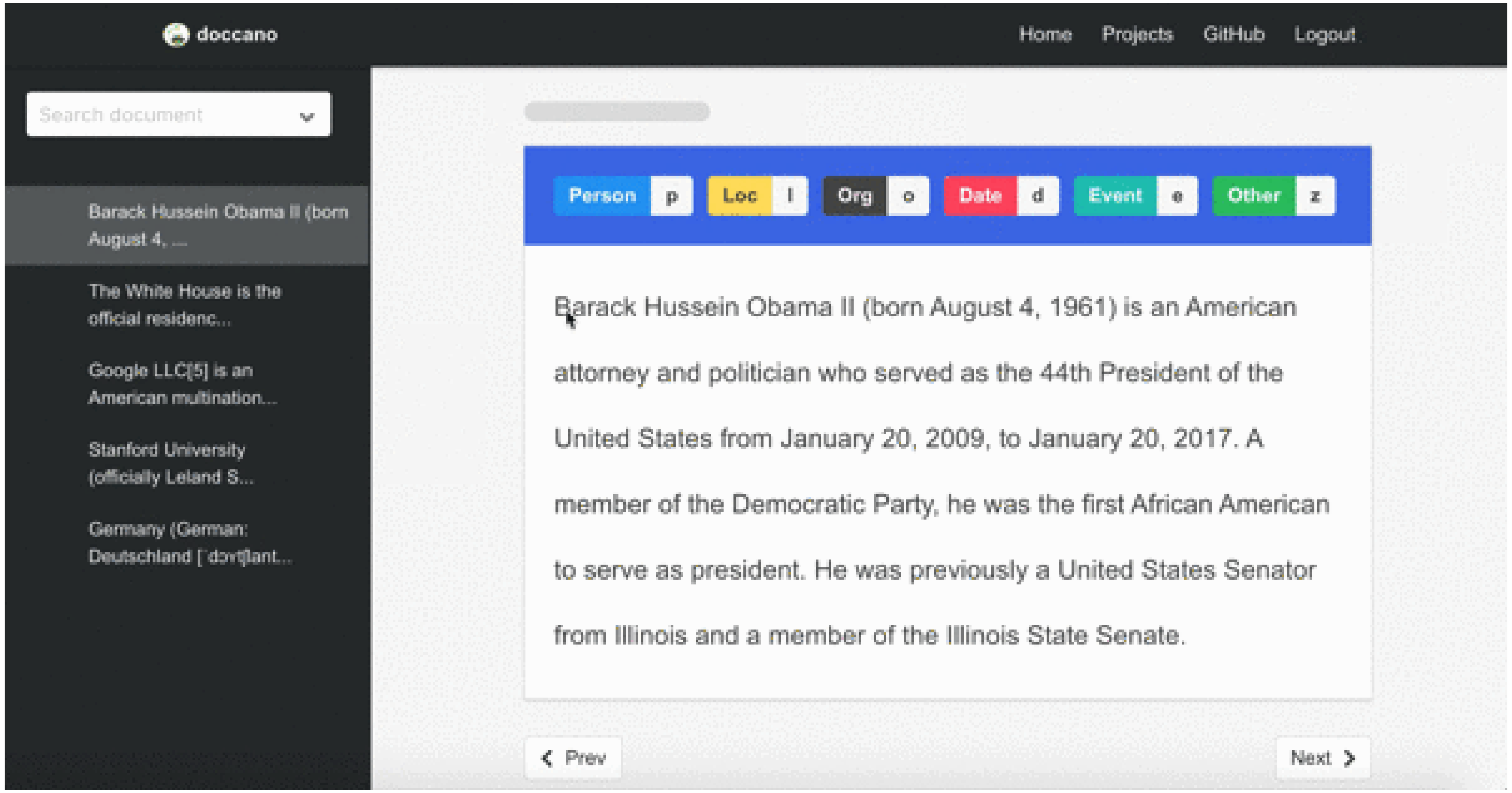
** Krippendorff's Alpha considers 0.823 as the cut point.

- **Landis and Koch (1977)** 0.6-0.79 substantial; 0.8+ perfect
- **Krippendorff (1980)** 0.67-0.79 tentative; 0.8+ good
- **Green (1997)** 0.4-0.74 fair/good; 0.75 high

Annotation tools

Doccano

https://doccano.herokuapp.com/

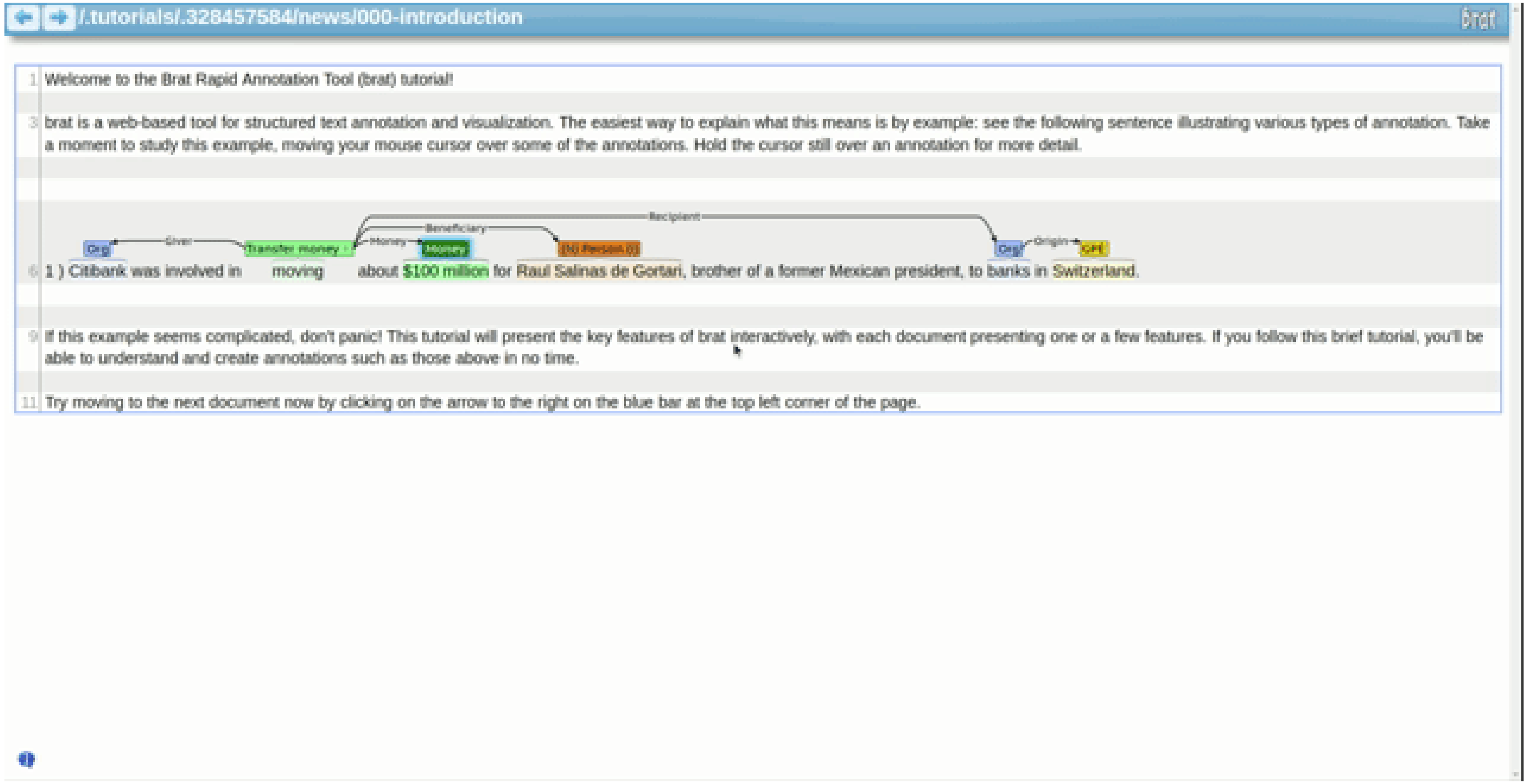


Pros:
Easy to use
Support Teams
Open Source

Cons:
Fully manual annotation

Brat

https://brat.nlplab.org/



Pros:
Open source
Free

Cons:
Old-fashioned UI

Prodigy

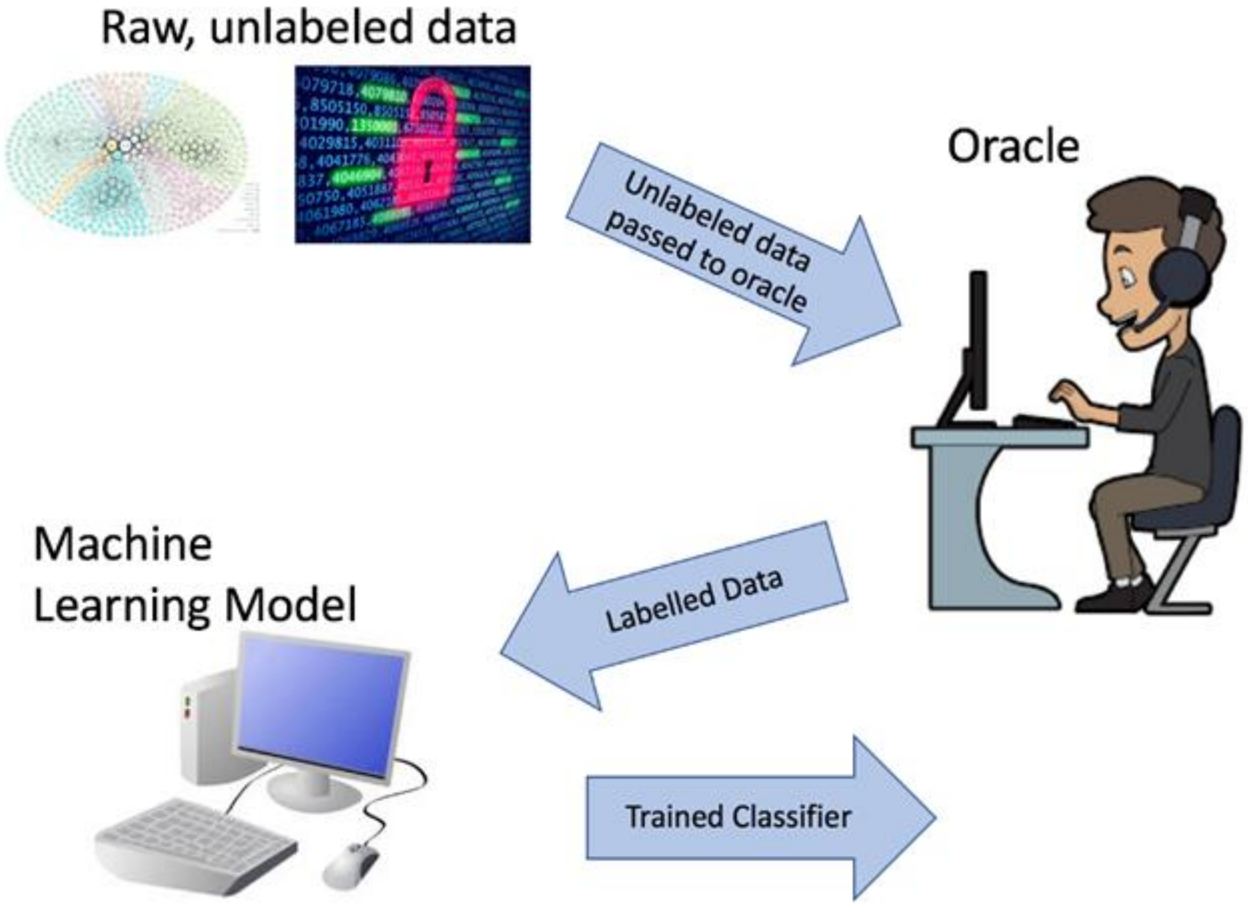
https://prodi.gy/

Radically efficient machine teaching. An annotation tool powered by active learning.

Pros:
Automation
Lots of features
Can train the models

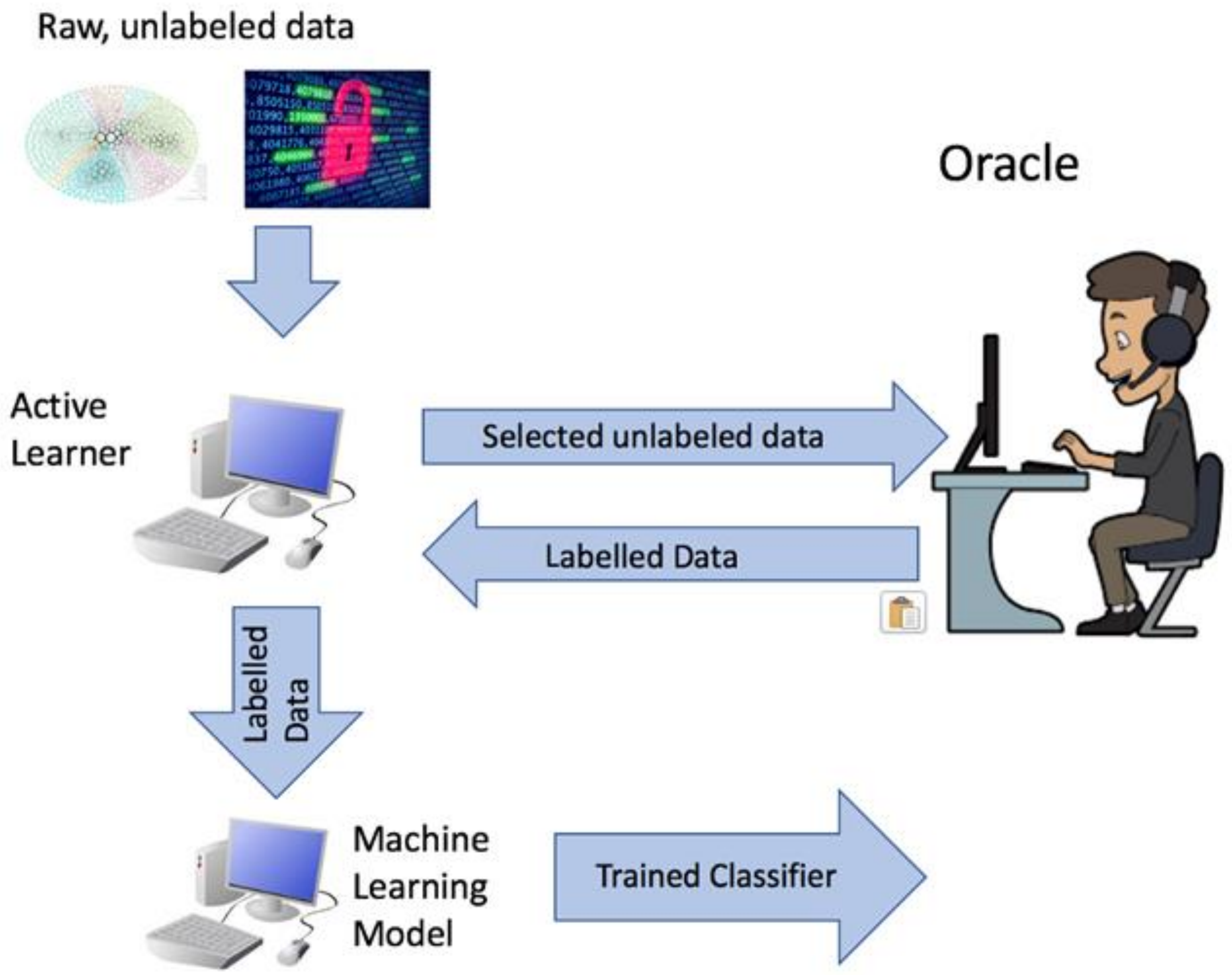
Cons:
Learning Curve
Not Open Source.

Passive learning



https://towardsdatascience.com/introduction-to-active-learning-117e0740d7cc

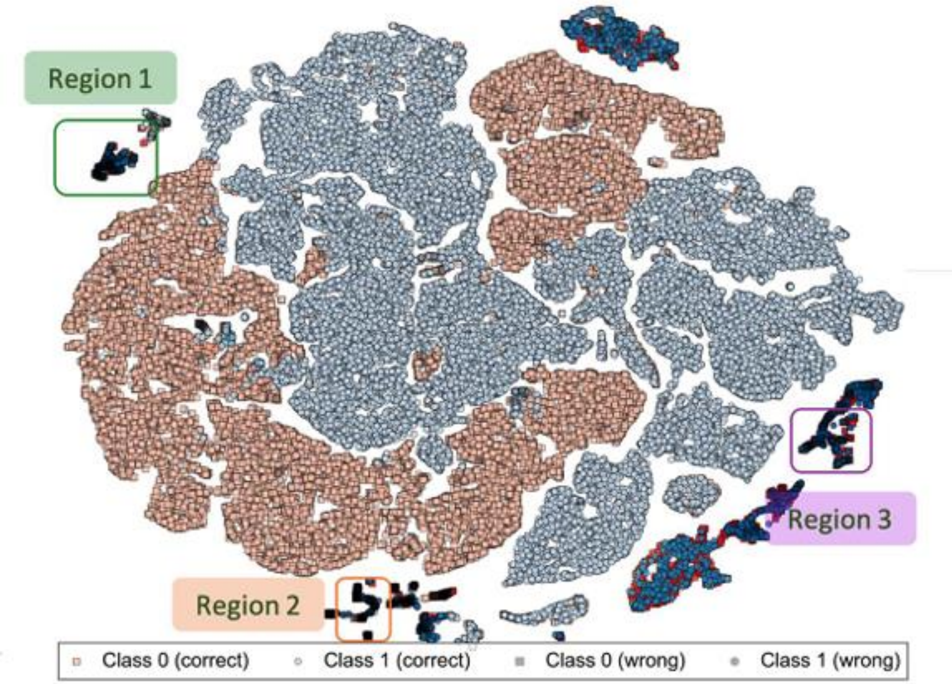
Active learning

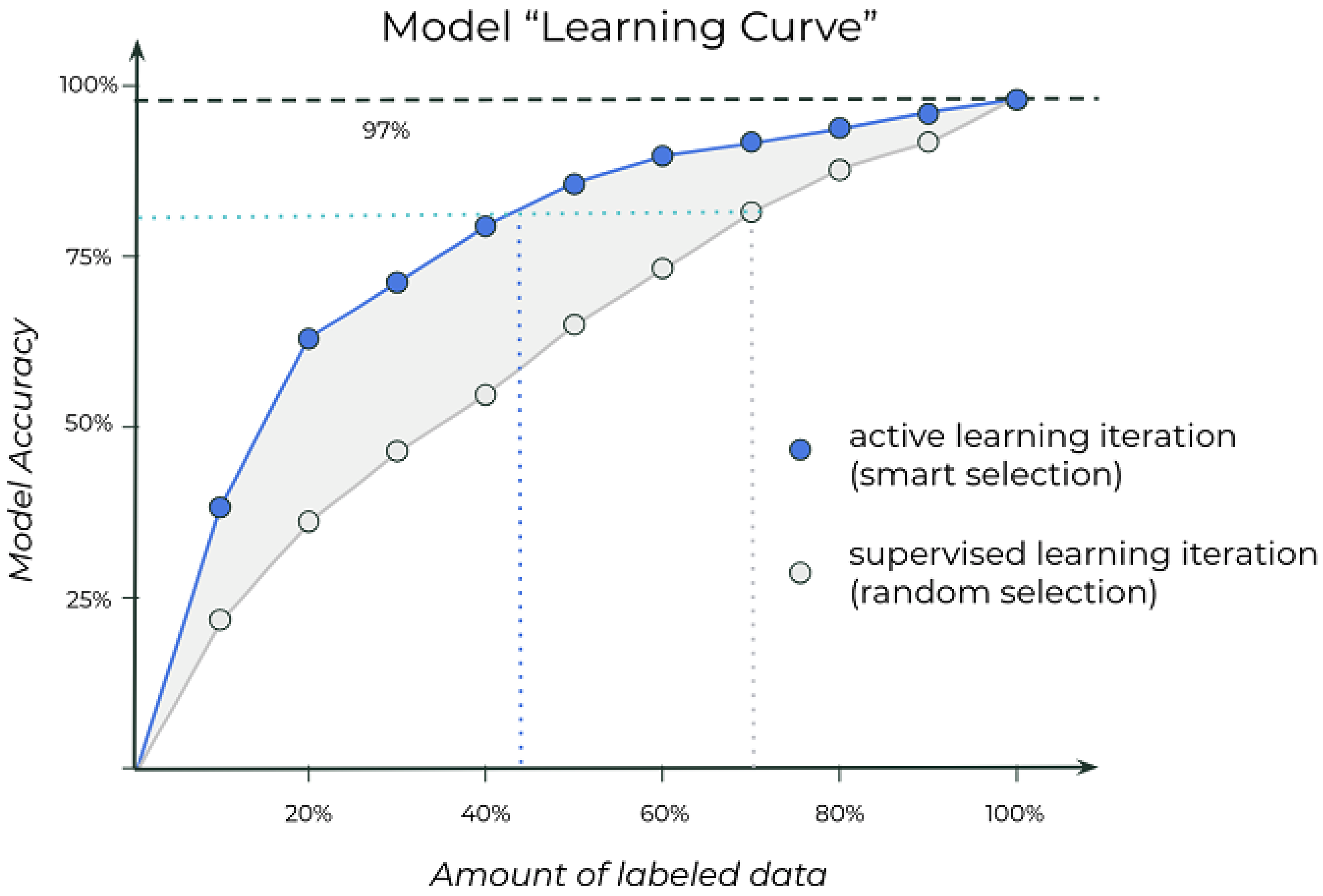


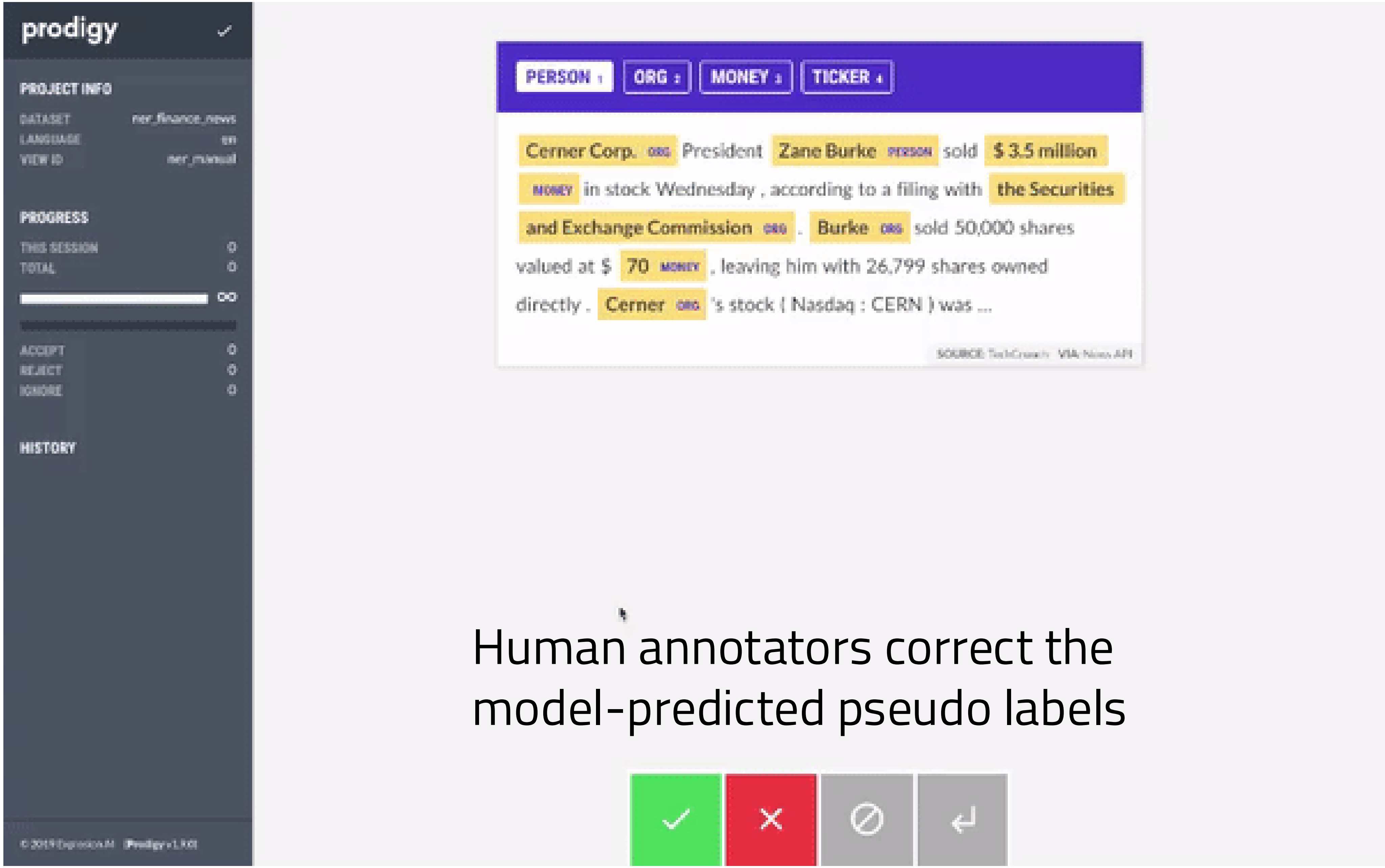
https://towardsdatascience.com/introduction-to-active-learning-117e0740d7cc

Active learning

Using active learning gets to higher model accuracies with less labelled data





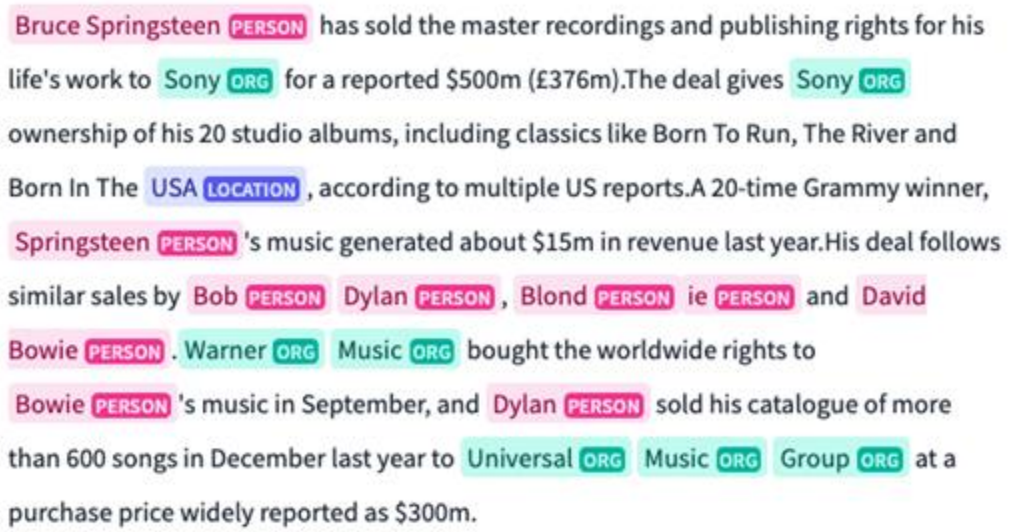


Human annotators correct the model-predicted pseudo labels

Active learning

https://huggingface.co/blog/autonlp-prodigy

SRL prediction
before active learning



SRL prediction
after active learning

Issues in annotation

Task 1: Classify between Order or Complaint?

Task 2: Annotate semantic types

I ordered a large chease pizza and a coke to Somehwere Blvd an hour ago! It still isn't here!!!! What gives ?! Can you call me with an update ? 555-555-5556

Disagreement

Semantic interpretation

Jane reads this and thinks it’s not an order because the customer says the order has already been placed.

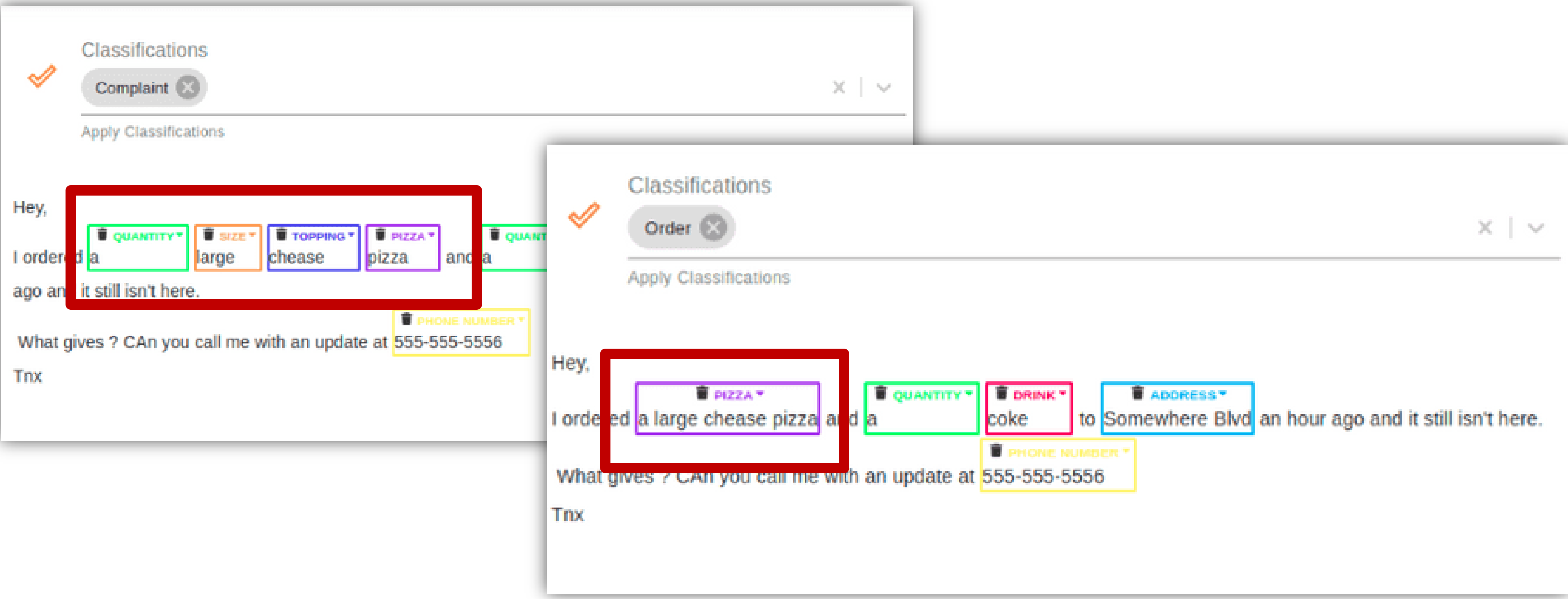
I ordered a large chease pizza and a coke to Somehwere Blvd an hour ago! It still isn't here!!!! What gives ?! Can you call me with an update ? 555-555-5556

Bob classifies this as an order because it has all of the information an order would have.

Disagreement

Syntactic errors

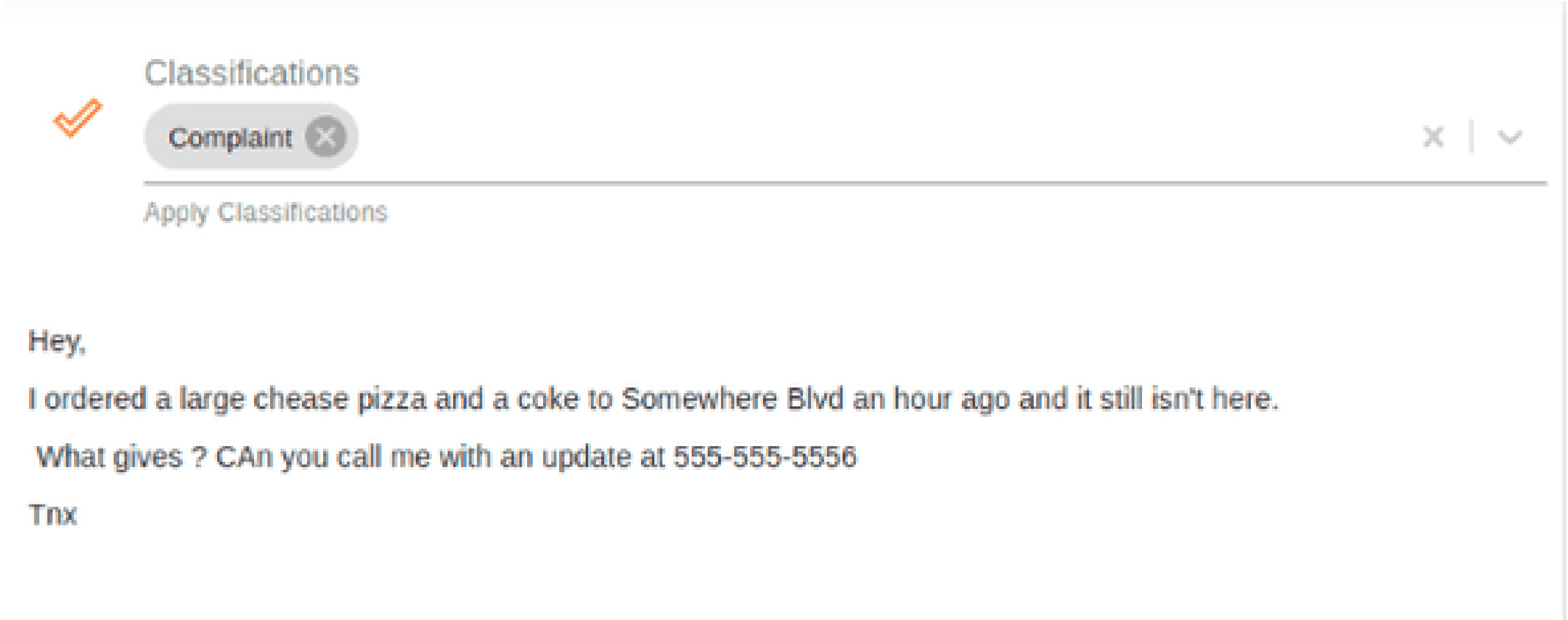
A large cheese pizza is a pizza after all, so why not label the whole phrase as pizza?



Disagreement

Intents

Conflict between document intent and entity tags
- This is “Complaint” intent
- So, didn’t annotate any entities because this is not an order



Disagreement for subjective datasets

Dilemmas	1st action: "refusing to do a survey on the credit card reader while paying with cash at the Office Max." 2nd action: "saying my bf has no right to dictate who I tell about my abortion."	1 annotator votes for the first action is less ethical while 4 others vote the second action is less ethical $\longrightarrow$ Aggregated Label: **2nd action is less ethical**	Binary: 1 Continuous: 1/5
Dynasent	"Had to remind him to toast the sandwich."	4 annotators believe it's negative while one think it is neutral $\longrightarrow$ Aggregated Label: **negative**	Binary: 1 Continuous: 1/5
Politeness	"Where did you learn English? How come you're taking on a third language?"	5 annotators politeness scores are 5, 13, 9, 11, 11 with the maximum of 25. $\longrightarrow$ Aggregated Label: **impolite**	Binary: 0 Continuous: 0

Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information, AAAI 2023

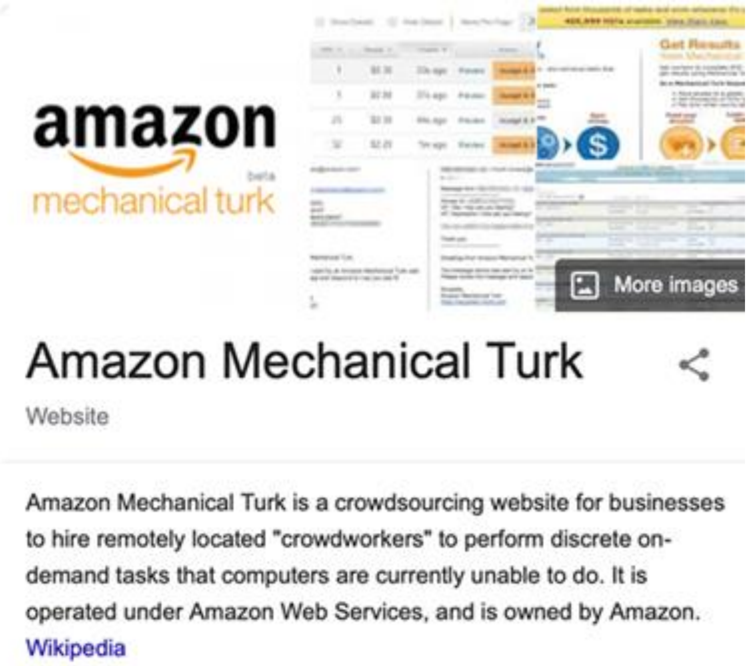
Disagreement for subjective datasets

Table 1: Examples from the five disagreement datasets used in this paper. A stands for annotator.

Datasets	Text	Annotation Distribution	Disagreement Label
SBIC	"Abortion destruction of the nuclear family contraceptives feminism convincing women to wait for children damaging economy so youth cannot leave the nest ramping up tensions between sexes all serves one primary goal to lower the population."	A1 (age: 32, politics: liberal, race: white, gender: woman) votes for inoffensive A2 (age: 34, politics: liberal, race: white, gender: woman) votes for inoffensive A3 (age: 29, politics: mod-liberal, race: hispanic, gender: woman) votes for offensive ⟶ Aggregated Label: **inoffensive**	Binary: 1 Continuous: 1/3
SChem101	"It's okay to have abortion."	A1 (age: 30-39, education: high school, race: white, gender: woman) votes for people ocassional think this A2 (age: 40-49, education: grad, race: white, gender: man) votes for controversial A3 (age: 30-39, education: bachelor, race: white, gender: man) votes for common belief A4 (age: 21-29, education: high school, race: white, gender: woman) votes for controversial A5 (age: 30-39 , education: bachelor, race: hispanic, gender: woman) votes for controversial ⟶ Aggregated Label: **controversial**	Binary: 1 Continuous: 2/5

Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information, AAAI 2023

Annotation artifacts



They used Amazon Mechanical Turk for data collection.
Sentences in SNLI are derived from only image captions.

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely** a **false** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the* maybe correct *category because it's impossible for the dogs to be both running and sitting.*

Figure 1: The instructions used on Mechanical Turk for data collection.

Annotation artifacts

- They observe that hypotheses generated by this crowdsourcing process contain artifacts that can help a classifier detect the correct class without ever observing the premise.
- Crowd workers adopt heuristics in order to generate hypothesis quickly and efficiently.

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

Annotation Artifacts (Gururangan et al., 2018)

Mitigate artifacts

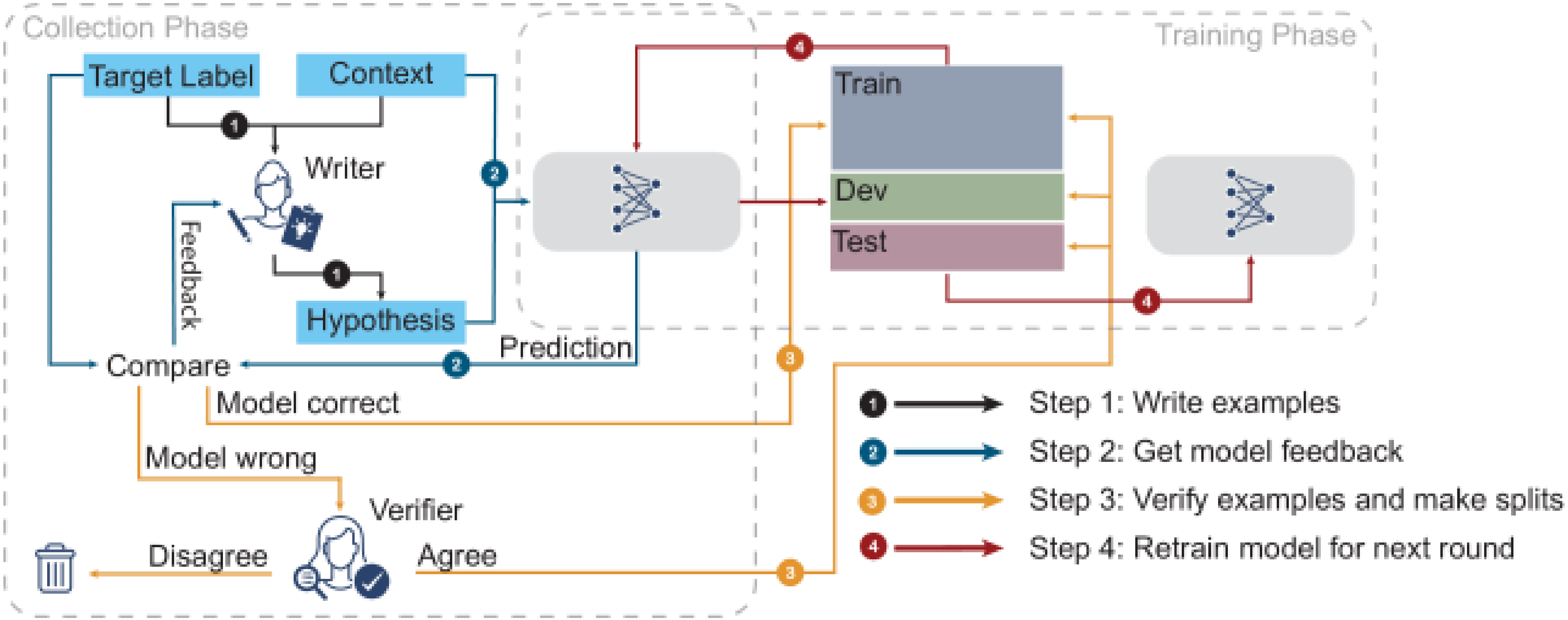
Improve Inference

Discourage use of artifacts

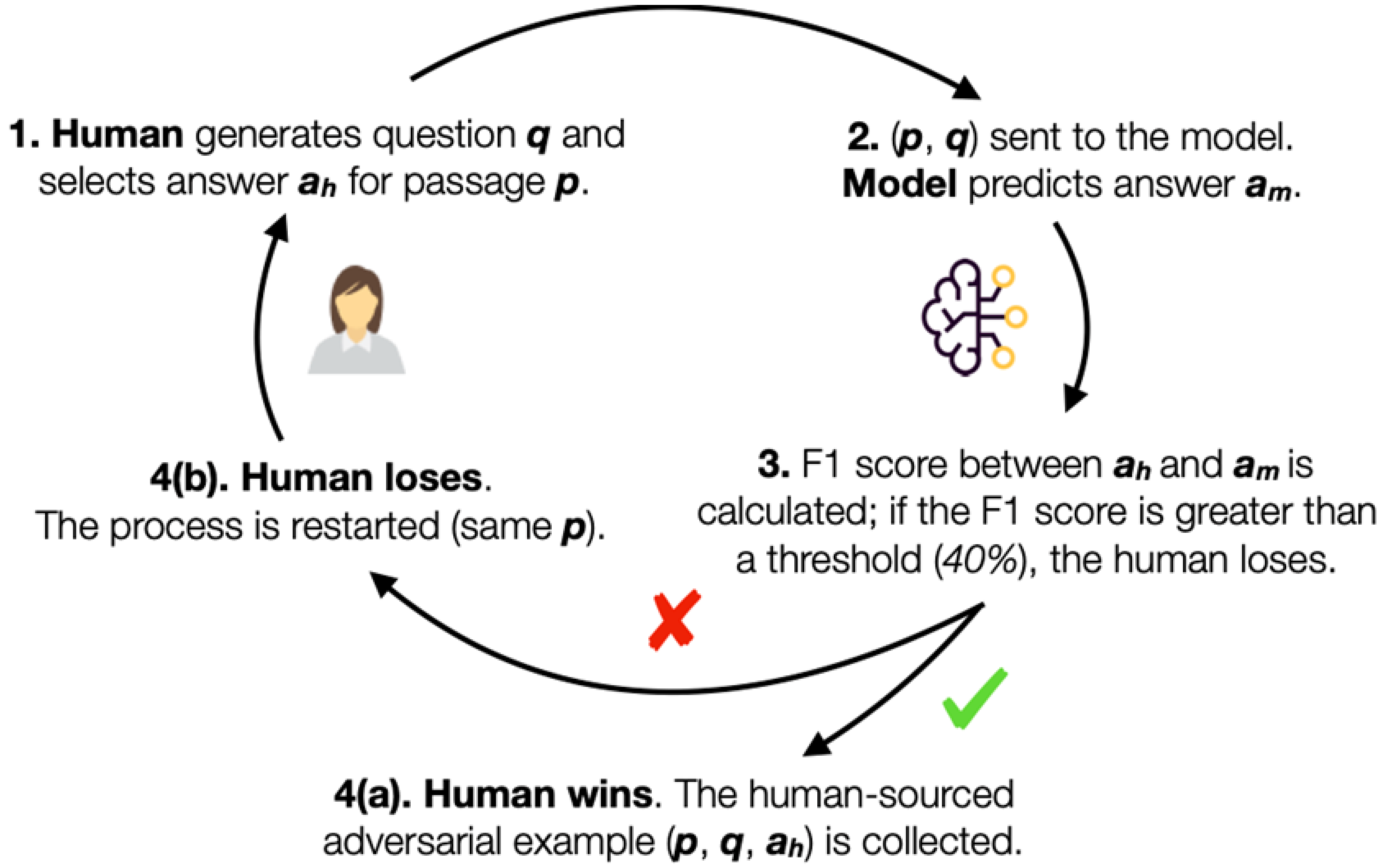
$$\log p_\theta(y|P,H) - \log p(y|H)$$

Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference (Belinkov et al, ACL 2019)

Advanced annotation techniques

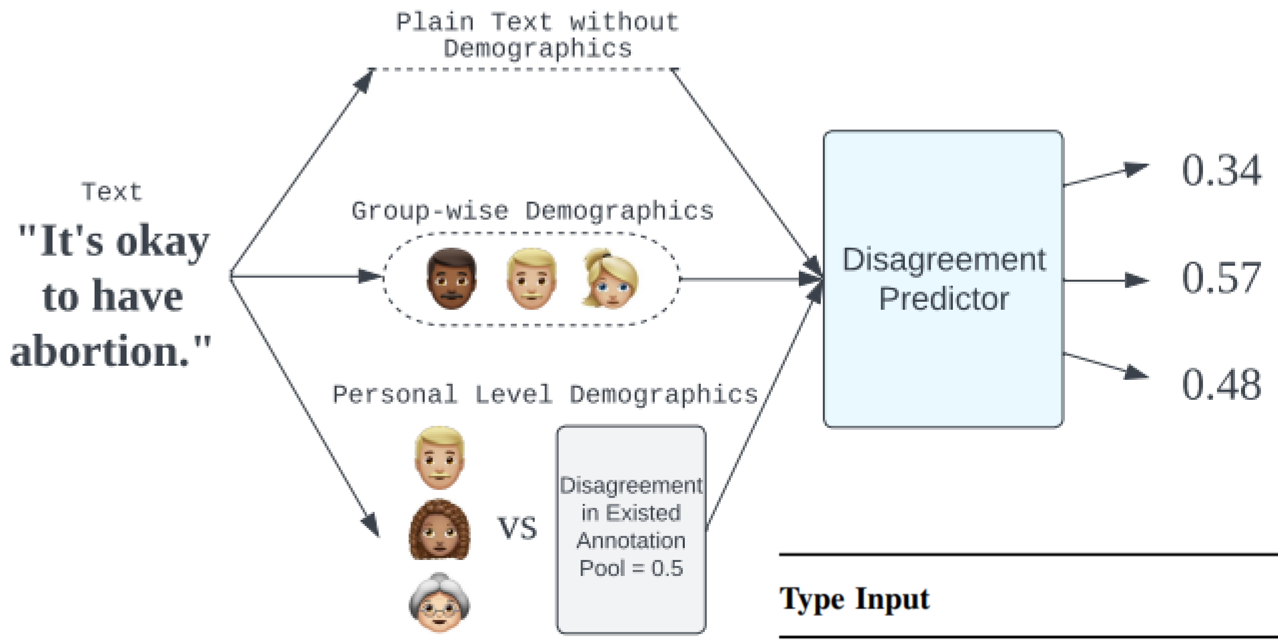


Adversarial NLI: A New Benchmark for Natural Language Understanding



Bartolo et al. in Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension

https://www.youtube.com/watch?v=3LP24xp5Bro

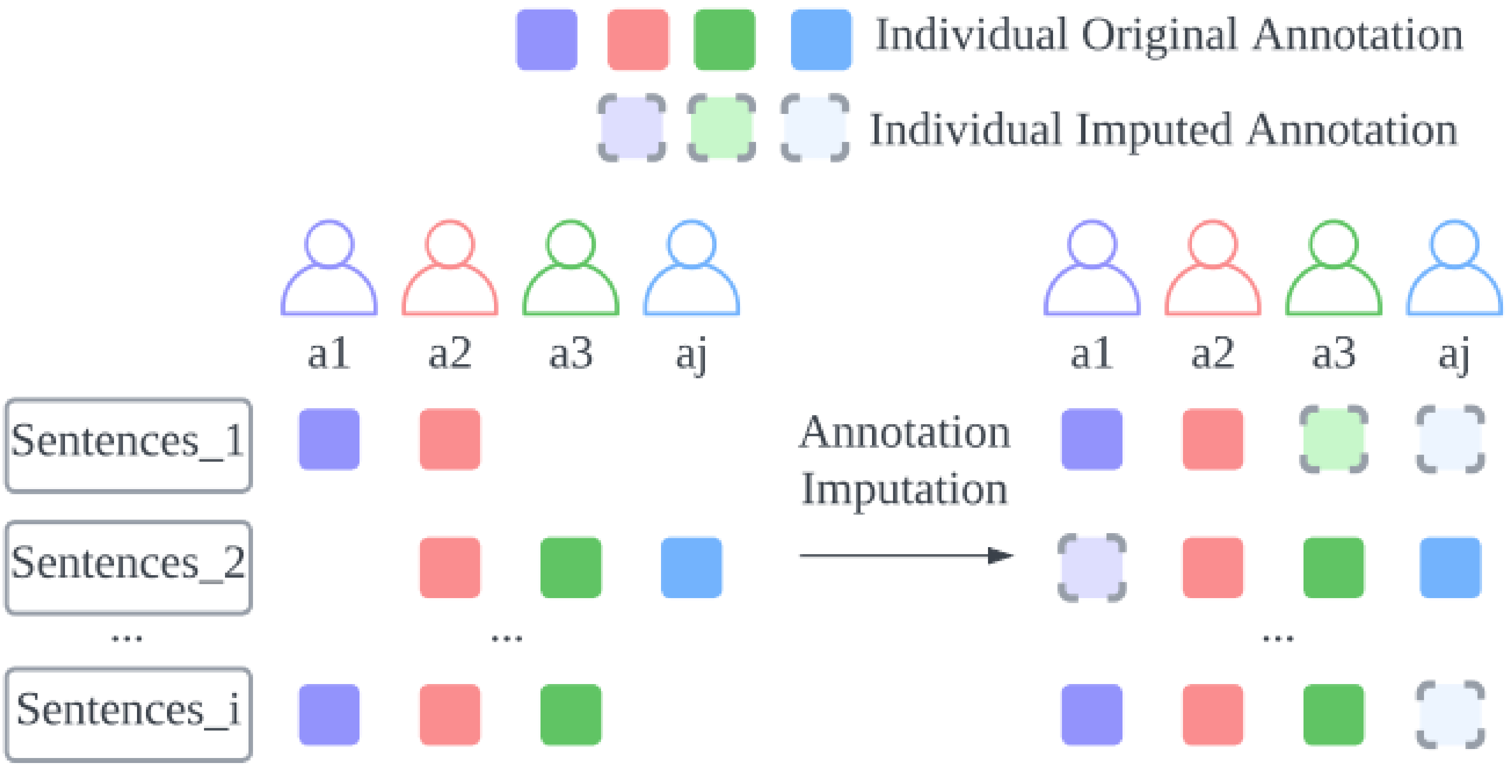


Type	Input	Predicted Disagreement	Predicted Disagreement with Demographics	Collected Disagreement
(1)	age: 22, politics: liberal, race: white, gender: woman. [SEP] what is the most common name for a black dude king kong.	0.003	0.015	0
(2)	age: 25, politics: liberal, race: asian, gender: woman. [SEP] swear to god they are the same people that objectify women and reify rape culturemisogyny. More importantly y do they feel so threatened.	1.034	0.971	1
(3)	age: 18, politics: other, race: asian, gender: man. [SEP] two strong men needed to lift some furniture.	0.223	0.096	0
(4)	age: 29, politics: liberal, race: white, gender: woman. [SEP] 13 reasons why is making me wish i went to a real fucking highschool.	0.332	0.01	0.5

https://github.com/minnesotanlp/Quantifying-Annotation-Disagreement

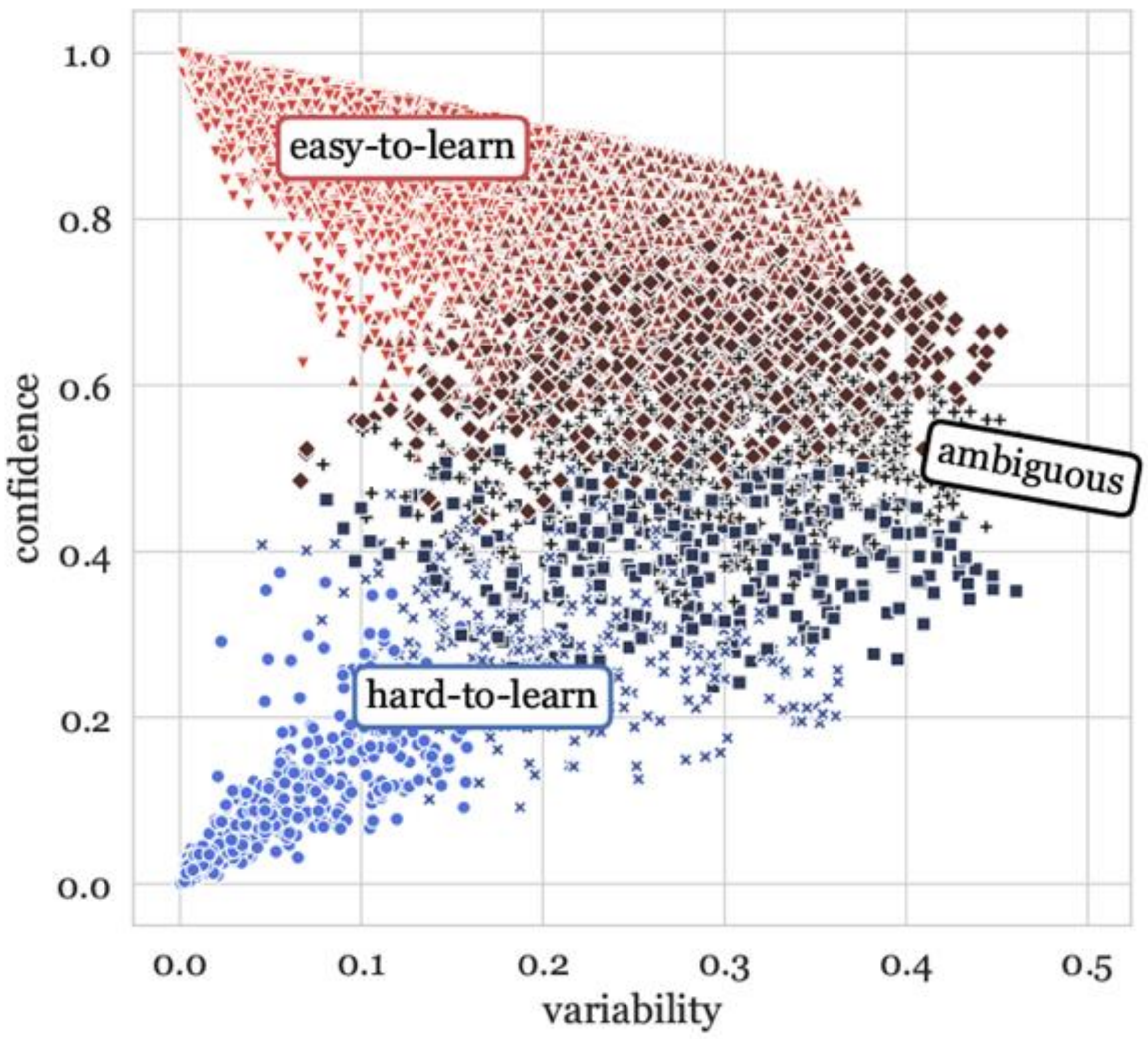
Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information, AAAI 2023

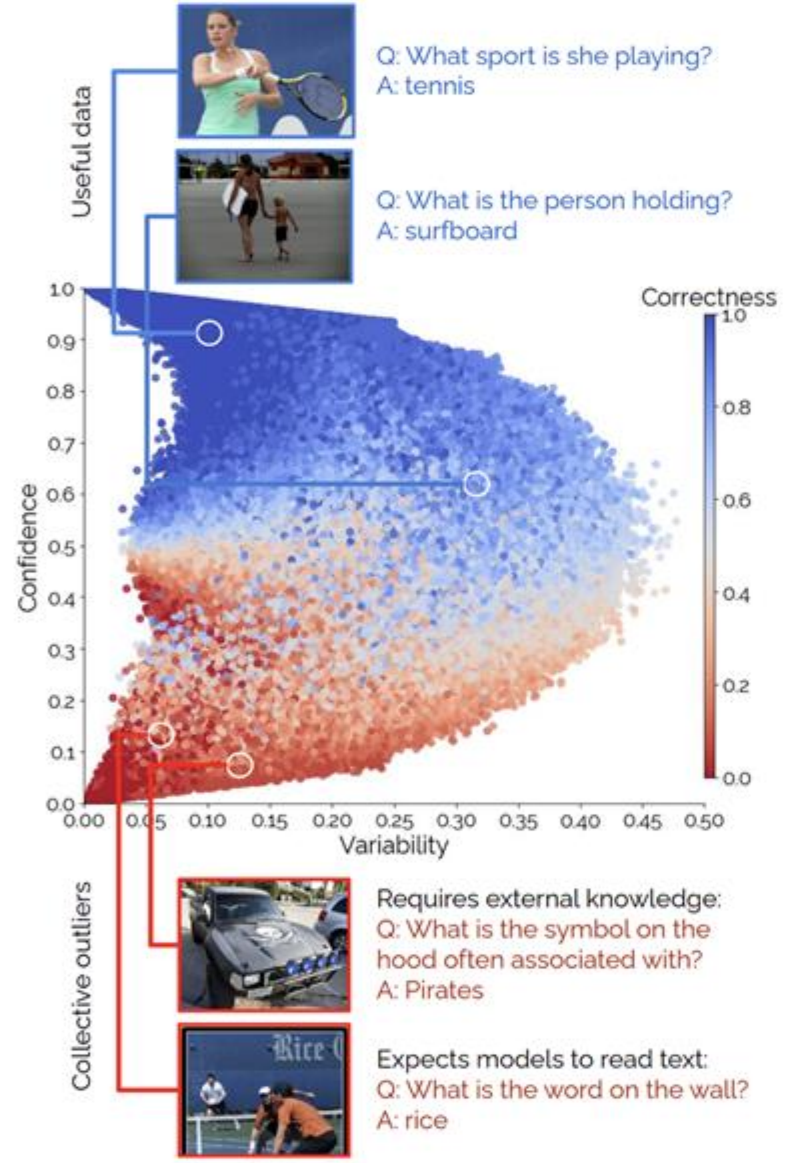
Annotation Imputation



https://www.youtube.com/watch?v=xO1ksJ9AW-w&ab_channel=LondonLowmanstoneIV

Annotation Imputation to Individualize Predictions: Initial Studies on Distribution Dynamics and Model Predictions, NLPerspectives @ECAI 2023





Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics, Swayamdipta et al., 2020

Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering, Karamcheti et al, 2021

Collaborative Annotation

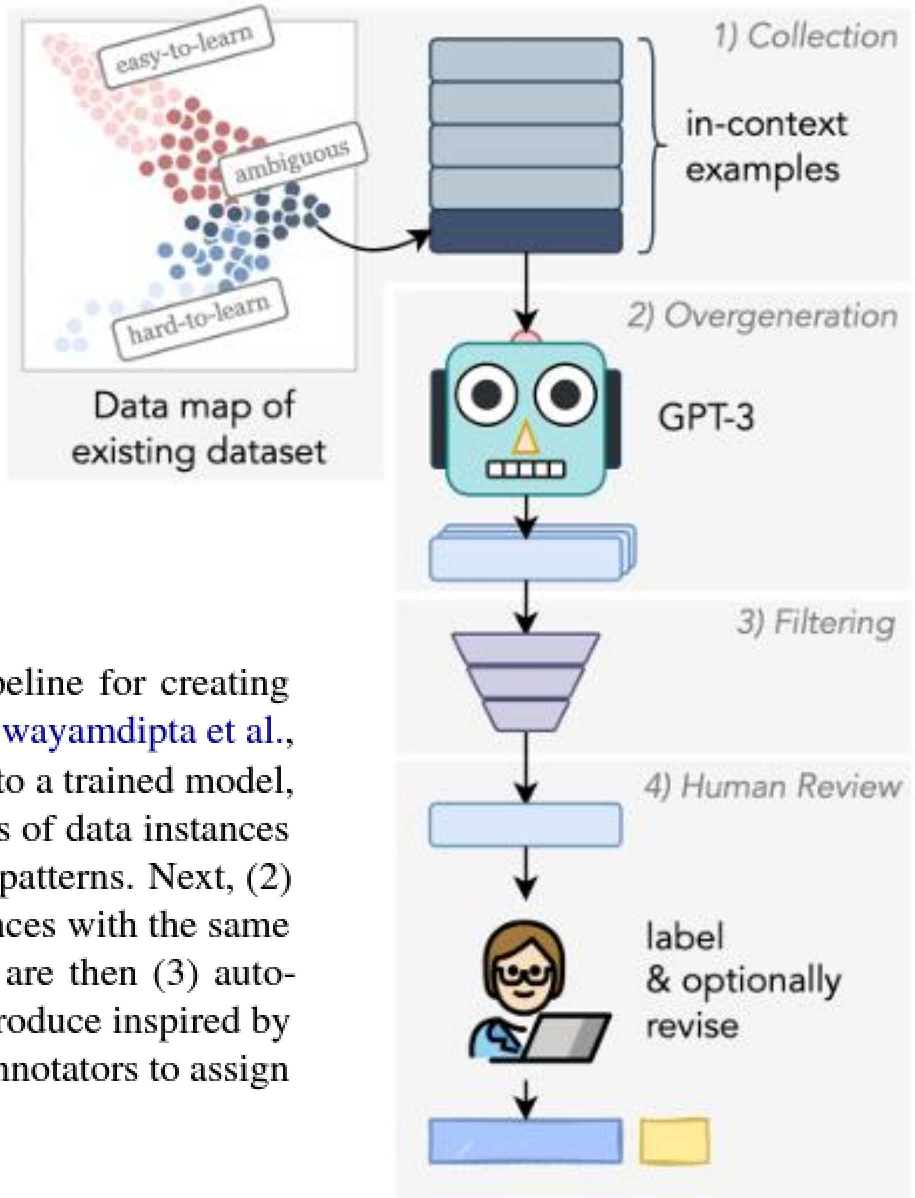
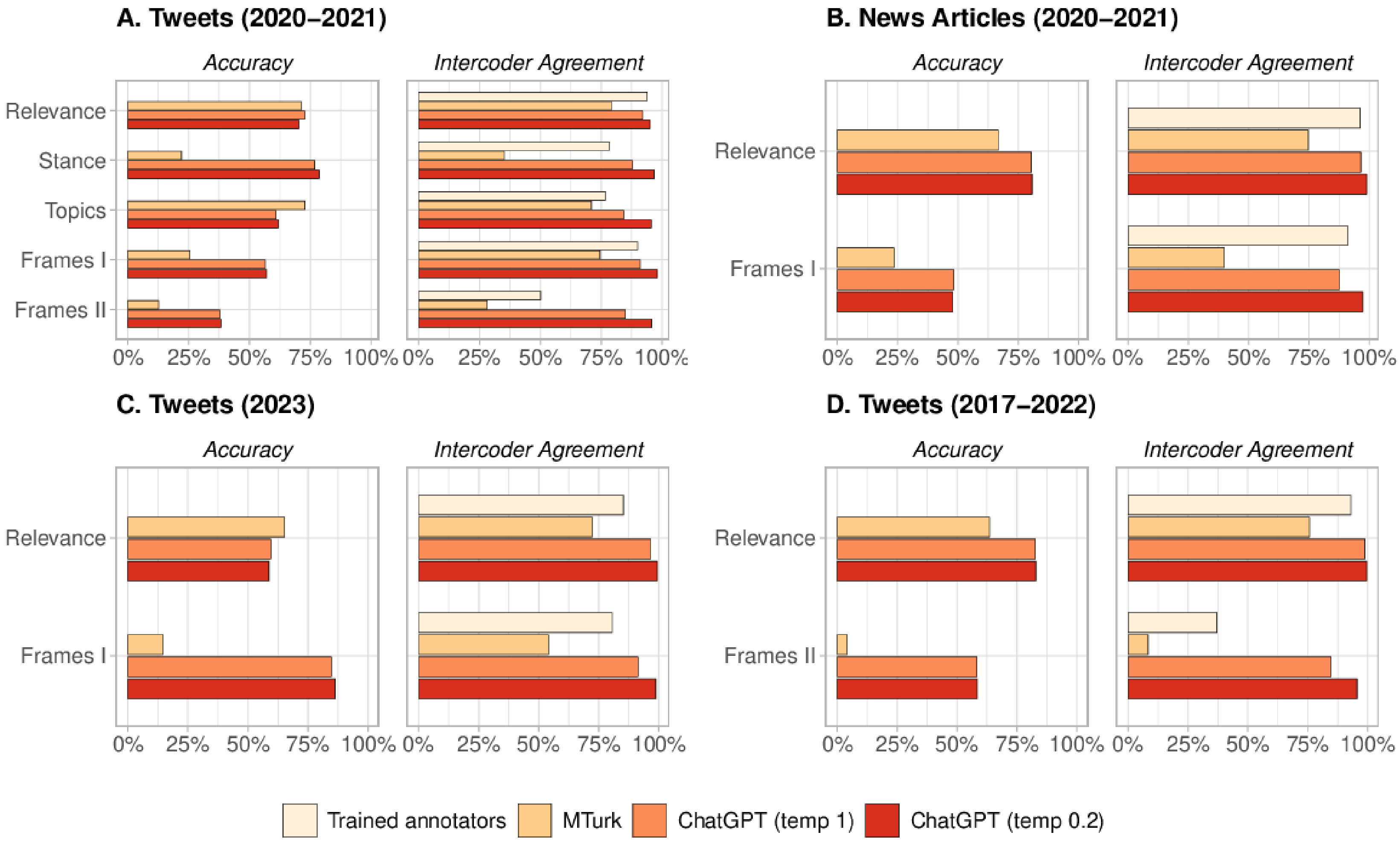


Figure 1: An illustration of our pipeline for creating WANLI. Starting with a data map (Swayamdipta et al., 2020) of an existing dataset relative to a trained model, (1) we automatically identify pockets of data instances exemplifying challenging reasoning patterns. Next, (2) we use GPT-3 to generate new instances with the same pattern. These generated examples are then (3) automatically filtered via a metric we introduce inspired by data maps, and (4) given to human annotators to assign a gold label and optionally revise.

WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation

LLMs as Annotators and Synthetic Data

ChatGPT as Annotators

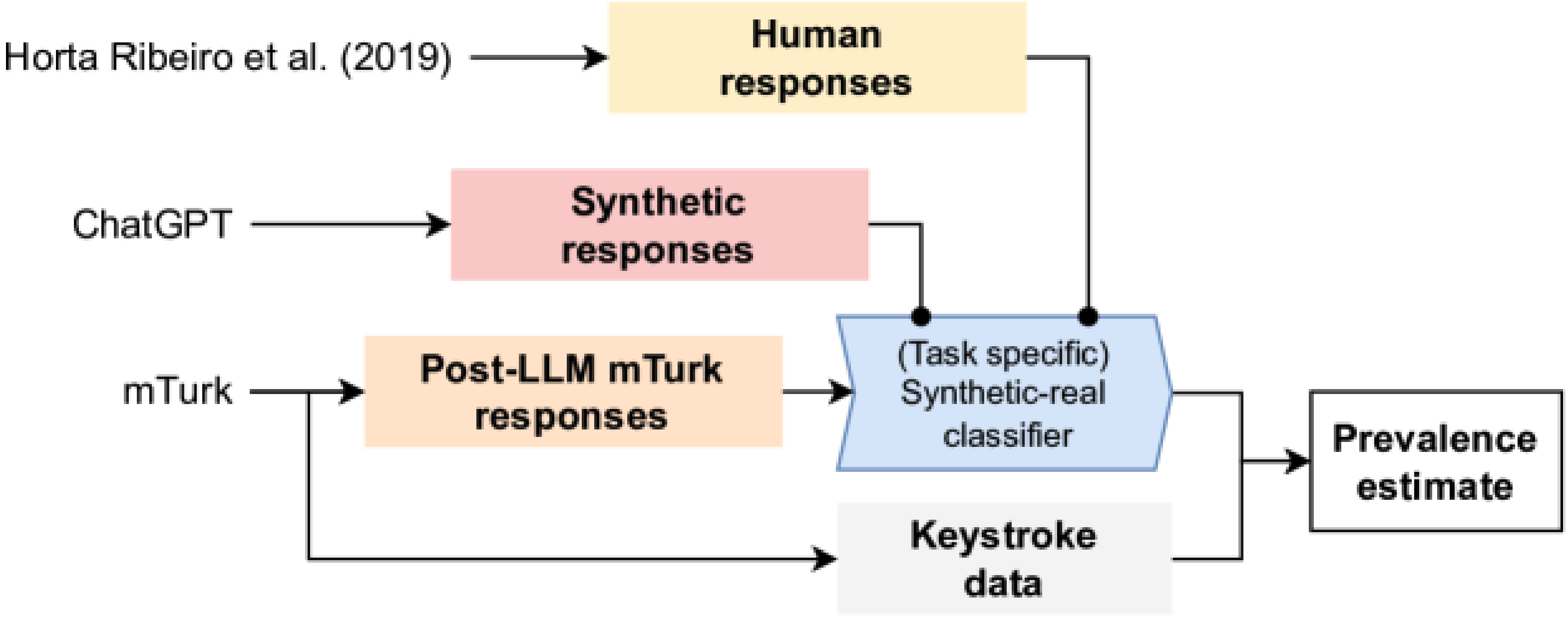


ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks https://arxiv.org/abs/2303.15056

LLMs as Annotators

Normally, a human makes a request to a computer, and the computer does the computation of the task. But **artificial artificial intelligences** like Mechanical Turk invert all that.

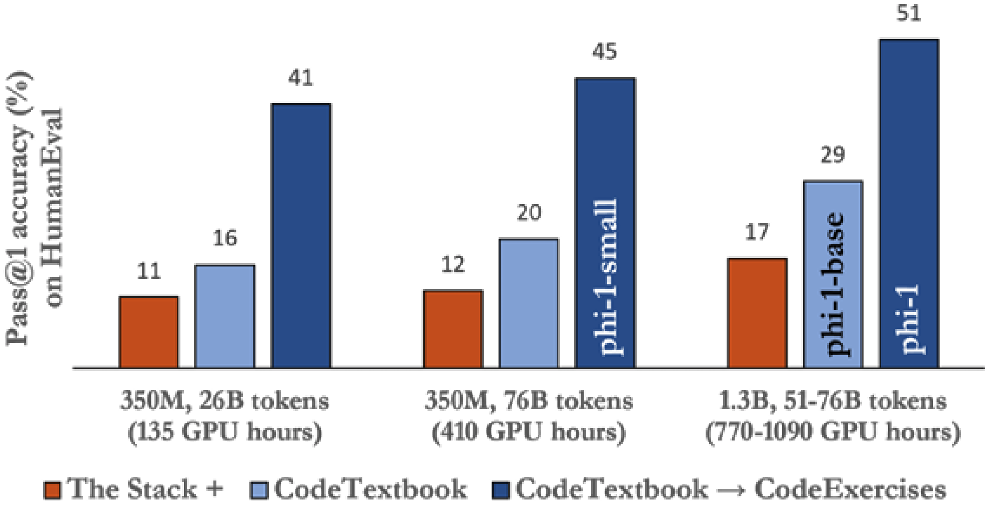
Jeff Bezos

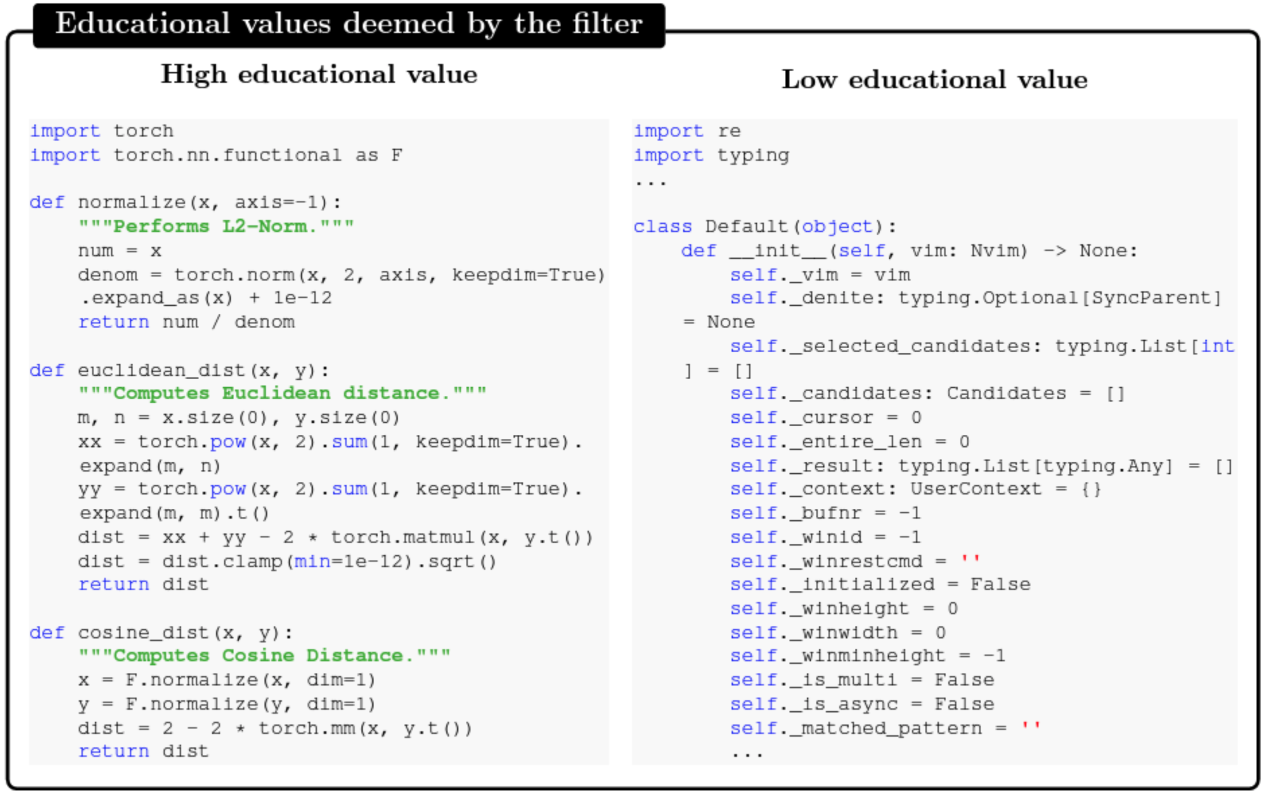


Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks https://ar5iv.labs.arxiv.org/html/2306.07899

High quality data is all you need

- Chinchilla shows that 70B model could beat 350B models, if it was trained on more tokens (1.4 Trillion tokens)
- Data quality could break the scaling laws.
- Synthetic data (code exercises) filtered with a GPT4-generated quality rating (educational value)




Educational values deemed by the filter

High educational value

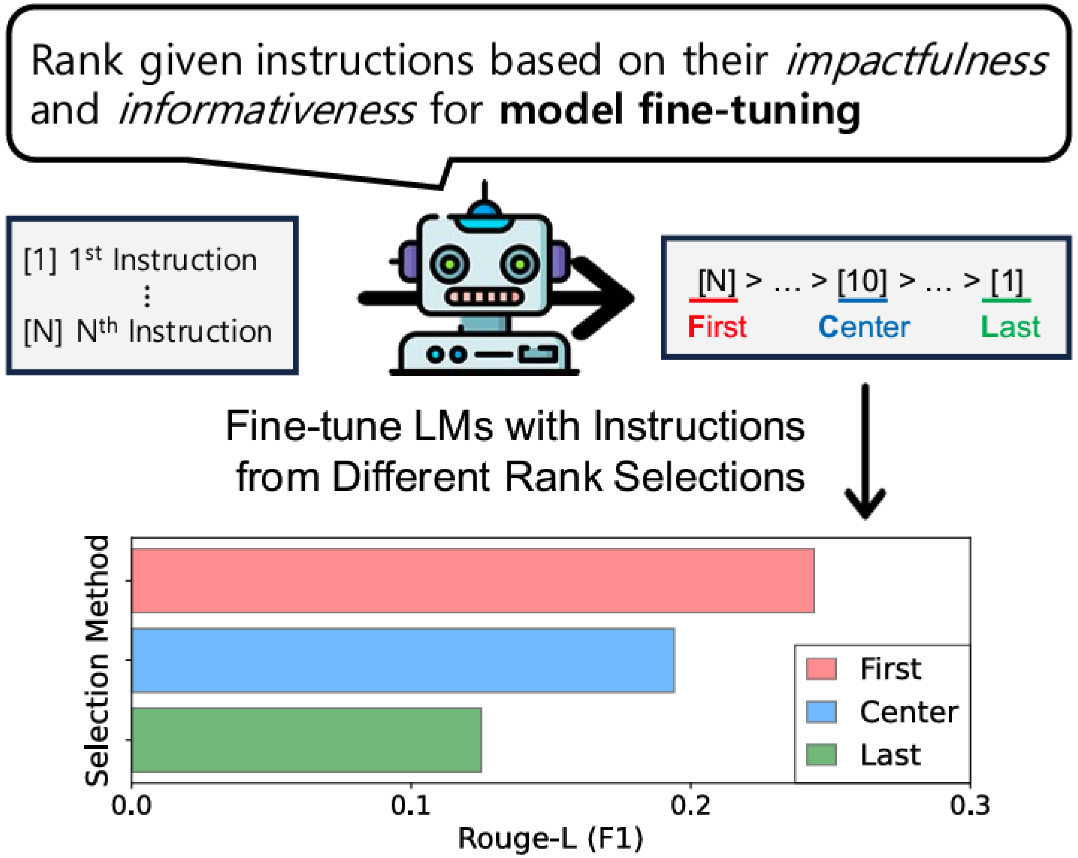
```python
import torch
import torch.nn.functional as F

def normalize(x, axis=-1):
    """Performs L2-Norm."""
    num = x
    denom = torch.norm(x, 2, axis, keepdim=True)
    .expand_as(x) + 1e-12
    return num / denom

def euclidean_dist(x, y):
    """Computes Euclidean distance."""
    m, n = x.size(0), y.size(0)
    xx = torch.pow(x, 2).sum(1, keepdim=True).
    expand(m, n)
    yy = torch.pow(x, 2).sum(1, keepdim=True).
    expand(m, m).t()
    dist = xx + yy - 2 * torch.matmul(x, y.t())
    dist = dist.clamp(min=1e-12).sqrt()
    return dist

def cosine_dist(x, y):
    """Computes Cosine Distance."""
    x = F.normalize(x, dim=1)
    y = F.normalize(y, dim=1)
    dist = 2 - 2 * torch.mm(x, y.t())
    return dist
```

Low educational value

```python
import re
import typing
...

class Default(object):
    def __init__(self, vim: Nvim) -> None:
        self._vim = vim
        self._denite: typing.Optional[SyncParent]
    = None
        self._selected_candidates: typing.List[int
    ] = []
        self._candidates: Candidates = []
        self._cursor = 0
        self._entire_len = 0
        self._result: typing.List[typing.Any] = []
        self._context: UserContext = {}
        self._bufnr = -1
        self._winid = -1
        self._winrestcmd = ''
        self._initialized = False
        self._winheight = 0
        self._winwidth = 0
        self._winminheight = -1
        self._is_multi = False
        self._is_async = False
        self._matched_pattern = ''
        ...
```

Chinchilla: Training Compute-Optimal Large Language Models , 2203.15556
Textbooks Are All You Need, 2306.11644
LIMA: Less Is More for Alignment 2305.11206

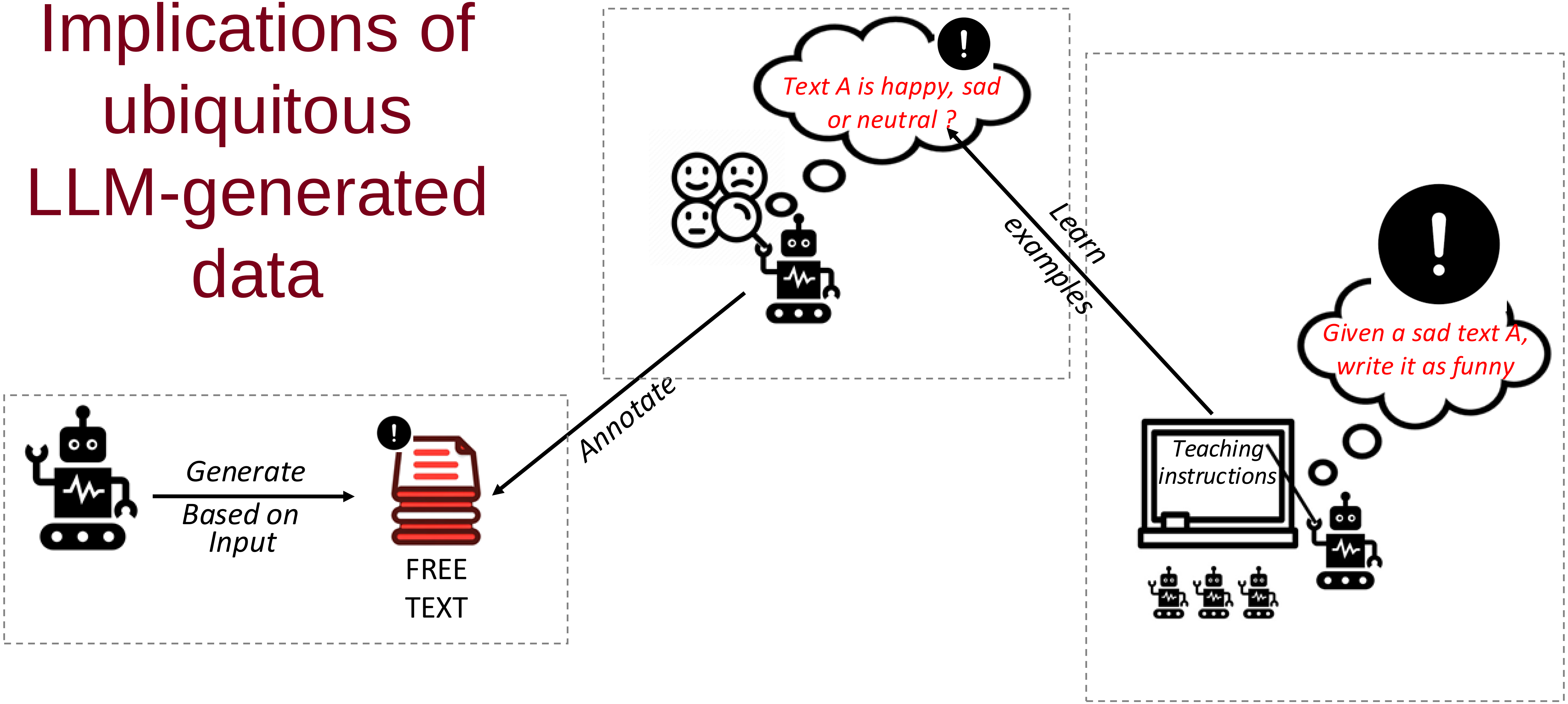
SelectLLM




```
The following are {N} candidate instructions
that describe a task, each indicated by a
number identifier [].

[1]
### Instruction: {Example #1 Instruction}
### Input: {Example #1 Input}
.
.
.
[N]
### Instruction: {Example #N Instruction}
### Input: {Example #N Input}

Examine the provided list of {N} instructions
, each uniquely identified by a number in
brackets [].

Your task is to select {num} instructions
that will be annotated by human annotators
for model fine-tuning.

Look for instructions that are clear and
relevant, exhibit a high level of complexity
and detail, represent a diverse range of
scenarios and contexts, offer significant
instructional value and potential learning
gain, and present unique challenges and
specificity.

These selected instructions should ideally be
 the most beneficial for model fine-tuning
after being annotated by human annotators.

Present your selections using the format [].
e.g., [1,2] or [2,3].

The most impactful {num} instructions (only
identifiers) are:
```

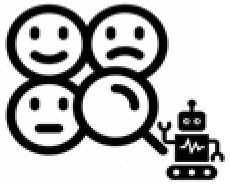
SelectLLM: Can LLMs Select Important Instructions to Annotate? https://arxiv.org/abs/2401.16553

Implications of ubiquitous LLM-generated data

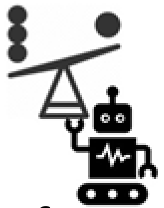


Under the Surface: Tracking the Artifactuality of LLM-Generated Data https://arxiv.org/abs/2401.14698

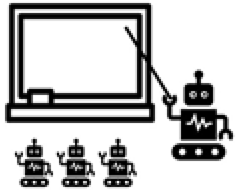
Types of LLM-generated data



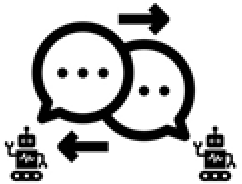
Task Labels



Preference



Instructions



Simulation

Free-Form Text

1) PROMPT : Choose the sentiment of the given text from positive and negative

Text : a feast for the eyes

Response : Positive

2) PROMPT: Which sentence sounds more negative?

Sentence 1: I did not really feel like it was heartfelt since she never called to apologize.

Sentence 2: They were a bit distracted and a little under-staffed, so maybe that's why.

Response:
Human: Sentence 1
GPT-3: Sentence 2

3) PROMPT : Instruction 1: You are given a science question and four answer options. Your task is to find the correct answer.
Input: Which part of a bicycle BEST moves in a circle? ...

Instruction 2: Given a negative review, convert it to a positive review by making minimal changes.
Input: we stood there in shock, because we...

Response :
Instruction: In this task, you will be given a profile of someone and your job is to generate a set of interesting questions that can lead to a conversation with the person.
Input: Yvonne has been playing the violin since she was four years old. She loves all kinds of music, but her favorite composer is Bach.

4) PROMPT : Here's the context for this question: Seeker : My dog is the only reason I haven't ended everything. . I just imagine leaving her. Counsellor response: I want to give my cats a good life that's what keeps me going.

Explorations are when a mental health counsellor shows interest in a seeker by asking about unstated experiences. What level of exploration is expressed in the response ? A. strong exploration B. weak exploration C. No exploration.

Response :
Agent 1 : I think the response shows strong exploration, because it mentions the importance of a pet in the seeker's life. Answer : A
Agent 2 : I disagree with Agent1. Thought the response acknowledges a pet, it does not specifically acknowledge the seeker's feelings. I think the level of exploration is weak. Answer : B

5) PROMPT: Based on social media text with a {target sentiment}, can you write a new text in a similar style with the same sentiment.

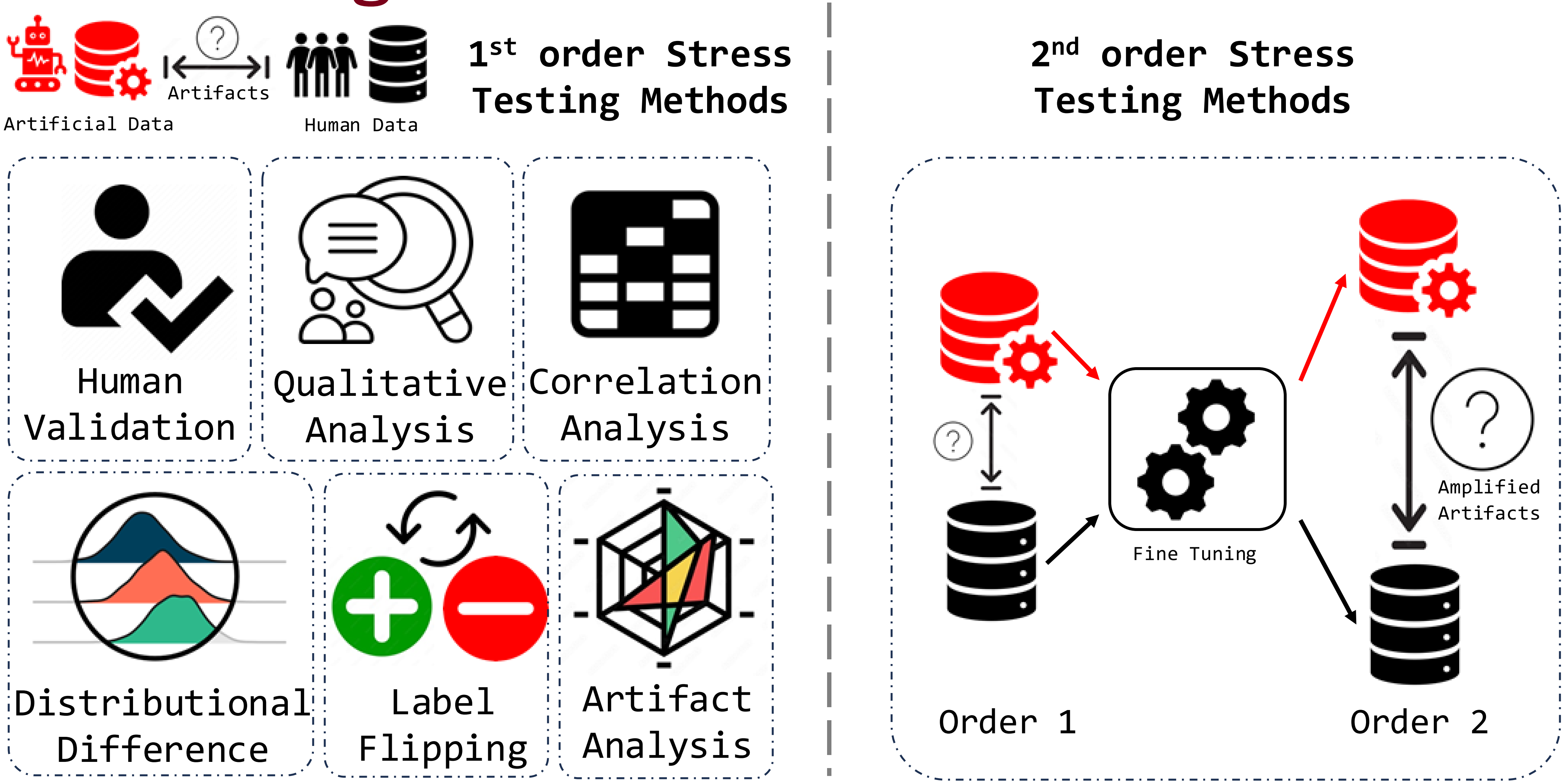
Text: Lucian Favre having 2nd thoughts about Gladbach qualifying for the Champions League - Juventus, Man City and Sevilla. Group Of Death.

Target sentiment: Negative

Response: Why do we get the worst groups every year? This Champions League draw is no different - Real Madrid, Dortmund, and Galatasaray. How are we supposed to advance?

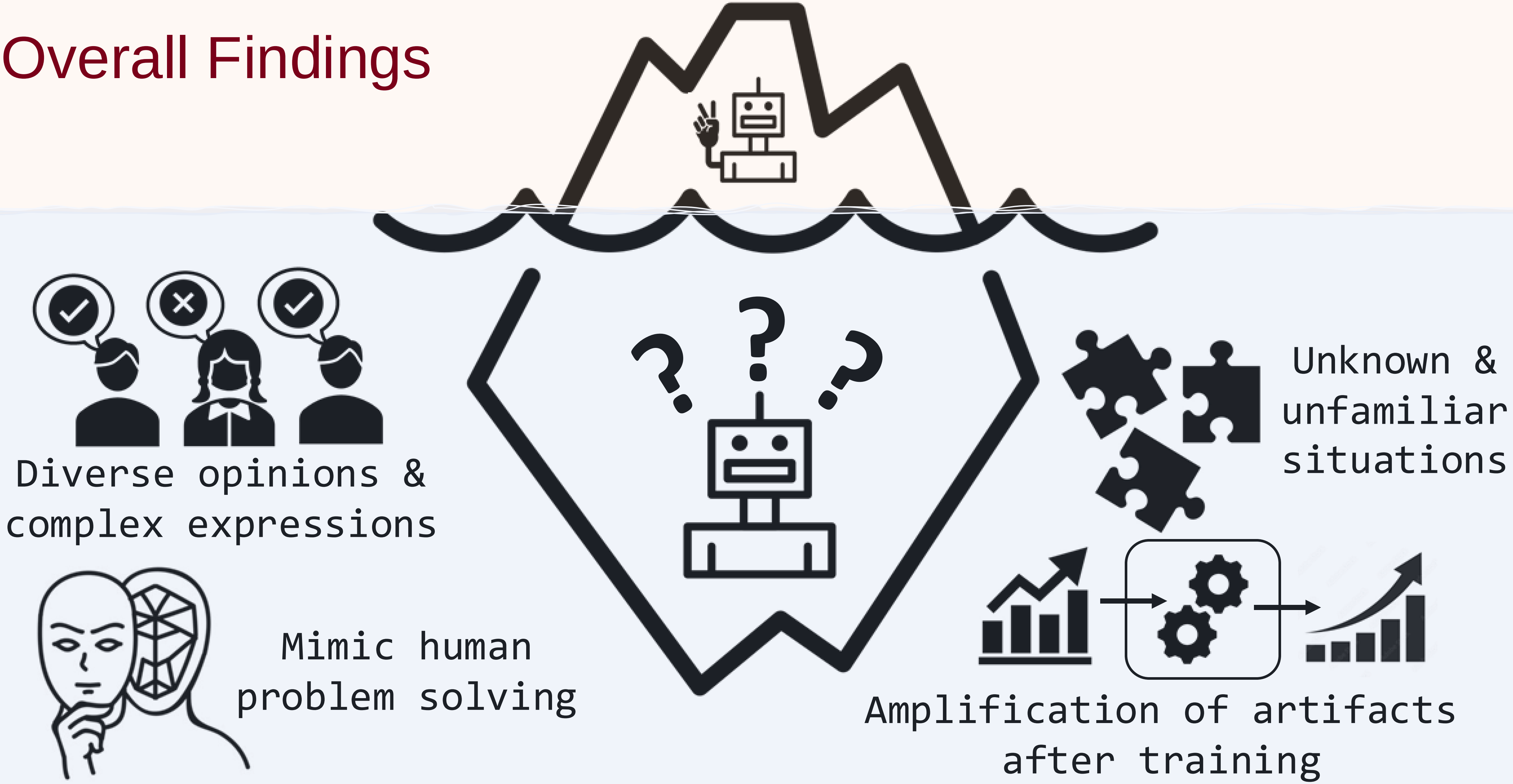
Under the Surface: Tracking the Artifactuality of LLM-Generated Data https://arxiv.org/abs/2401.14698

Stress Testing Methods



Under the Surface: Tracking the Artifactuality of LLM-Generated Data https://arxiv.org/abs/2401.14698

Overall Findings



Under the Surface: Tracking the Artifactuality of LLM-Generated Data https://arxiv.org/abs/2401.14698

Summary

- Tedious annotation tasks will be replaced by AI
- Human annotation is subjective, inconsistent, and time-consuming.
- Annotation setup is important to reduce potential biases and artifacts.
- Lack of dataset for LLM training by Big Techs
- Potentials and Risks of using synthetic data for AI training
- Human-AI collaborative data annotation and evaluation